# UNIVERSIDAD COMPLUTENSE DE MADRID

## FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

## DEPARTAMENTO DE ECONOMÍA FINANCIERA Y CONTABILIDAD I



# TESIS DOCTORAL

## Análisis del riesgo de caída de cartera en seguros: metodologías de "inteligencia artificial" vs "modelos lineales generalizados"

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR


**María de Lourdes Gutiérrez Cordero**

DIRECTORAS

**María Jesús Segovia Vargas**
**Susana Blanco García**


**Madrid, 2017**

# UNIVERSIDAD COMPLUTENSE DE MADRID

**FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES**

**DEPARTAMENTO DE ECONOMÍA FINANCIERA Y CONTABILIDAD I**


# Análisis del Riesgo de Caída de Cartera en Seguros: Metodologías de "Inteligencia Artificial" vs "Modelos Lineales Generalizados"

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR


**María de Lourdes Gutiérrez Cordero**

DIRECTORES

**María Jesús Segovia Vargas**

**Susana Blanco García**


MADRID, SEPTIEMBRE 2015

**DEPARTAMENTO DE ECONOMIA FINANCIERA Y CONTABILIDAD I**

**FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES**

# Análisis del Riesgo de Caída de Cartera en Seguros: Metodologías de "Inteligencia Artificial" vs "Modelos Lineales Generalizados"

TESIS DOCTORAL

**AUTOR:**

María de Lourdes Gutiérrez Cordero

**DIRECTORES:**

María Jesús Segovia Vargas

Susana Blanco García

MADRID, SEPTIEMBRE 2015

**DEDICATORIA**

Dedico esta tesis:

A mis padres, por los valores, principios, confianza y todo el amor que me han dado para lograr mis metas…

A mi hermano, el ser más noble que nunca dejará de contar conmigo…

A Memo, al confidente que más admiro y amo, es un privilegio caminar a tu lado…

A mis directoras de tesis, sin su apoyo y dedicación esto no hubiera sido posible…

A mis amigos de México, su cariño nunca se dejó de sentir a pesar de la distancia…

A toda la gente maravillosa de España, por hacernos sentir en familia tan lejitos…

*La derrota tiene algo positivo, nunca es definitiva. En cambio la victoria tiene algo negativo, jamás es definitiva.*

*–José Saramago-*

**INDICE**

**INDICE ..... 4**

**RESUMEN ..... 8**

**ABSTRACT ..... 12**

**CAPITULO 1: PROYECTO SOLVENCIA II ..... 15**

1.1 Introducción ..... 15

1.2 Antecedentes ..... 17

1.3 El Proyecto Solvencia II ..... 20

1.4 Marco Regulatorio ..... 25

1.5 Los 3 Pilares que la componen ..... 29

1.5.1 Pilar I – Cuantitativo ..... 30

1.5.1.1 Fórmula Estándar ..... 32

1.5.1.2 Modelos Internos ..... 34

1.5.2 Pilar II – Cualitativo ..... 36

1.5.3 Pilar III – Disciplina del Mercado ..... 38

**CAPITULO 2: RIESGO DE CAÍDA DE CARTERA ..... 40**

2.1. Introducción ..... 40

2.2. Riesgo de Caída de Cartera ..... 42

2.3. Estadísticas y Causas del Riesgo de Caída de Cartera ..... 45

**CAPITULO 3: TRATAMIENTO DE LA INFORMACIÓN ..... 49**

3.1. Introducción ..... 49

3.2. Contexto ..... 50

3.3. Muestra ..... 53

3.4. Variables Utilizadas ..... 55

3.5. Discretización de Variables ..... 80

**CAPITULO 4: APLICACIÓN DE LAS TÉCNICAS DE INTELIGENCIA ARTIFICIAL ..... 84**

4.1. Introducción ..... 84

4.2. Inteligencia Artificial ..... 88

4.2.1. Técnica de Arboles de Decisión ..... 89

4.2.2. Teoría de Rough Set ..... 96

4.3. Aplicación Empírica de la Técnica de Arboles de Decisión ..... 105

4.3.1. Resumen de Validación de Resultados bajo el Algoritmo C4.5 ..... 106

4.3.2. Análisis de las Principales Ramas ..... 107

4.3.3. Análisis de los Principales Patrones de las Pólizas Recientes ..... 110

4.3.3.1. Arboles de Decisión correspondientes a la CLASE 1=Cancelación ..... 111

4.3.3.2. Arboles de Decisión correspondientes a la CLASE 0=Retención ..... 116

4.3.4. Principales Reglas de Decisión de las Pólizas Recientes ..... 121

4.3.4.1. Reglas de Decisión correspondientes a la CLASE 1=Cancelación ..... 122

4.3.4.2. Reglas de Decisión correspondientes a la CLASE 0=Retención ..... 124

4.3.5. Principales Resultados Obtenidos bajo Arboles de Decisión ..... 127

4.4. Aplicación Empírica de la Técnica de Rough Set ..... 129

4.4.1. Resumen de Validación de Resultados bajo Rough Set ..... 130

4.4.2. Resumen de las Principales Reglas ..... 131

4.4.2.1. Reglas para la CLASE 1=Cancelación ..... 132

4.4.2.2. Reglas para la CLASE 0=Retención ..... 133

4.4.3. Resumen de las Principales Variables ..... 134

4.4.4. Principales Resultados Obtenidos bajo Rough Set ..... 135

**CAPITULO 5: APLICACIÓN DE LA METODOLOGÍA DE MODELOS LINEALES GENRALIZADOS ..... 138**

5.1. Introducción ..... 138

5.2. Modelos Lineales Generalizados ..... 142

5.2.1. Marco Teórico ..... 143

5.2.2. Estructura y Parámetros ..... 144

5.2.3. Componentes ..... 146

5.2.4. Familia Exponencial ..... 148

5.2.5. Función Enlace ..... 149

5.2.6. Offset ..... 150

5.2.7. Estimación ..... 151

5.2.8. Estructuras de Modelos Comunes ..... 152

5.2.9. Validación del Modelo ..... 153

5.2.10. Sobredispersión ..... 158

5.2.11. Residuos ..... 159

5.3. Los GLM en la Práctica ..... 162

5.3.1. Análisis Preliminar ..... 163

5.3.2.- Iteración del Modelo ..... 166

5.3.3. Depuración del Modelo ..... 175

5.3.4. Interpretación de Resultados ..... 176

5.3.5. Ventajas y Limitaciones ..... 177

5.4. Aplicación Empírica ..... 179

5.4.1 Análisis Preliminares ..... 180

5.4.1.1. Análisis Univariante ..... 181

5.4.1.2. Análisis Bivariante ..... 182

5.4.2. Aplicación del Modelo ..... 188

5.4.2.1. Análisis Factorial ..... 188

5.4.2.2. Elección del Modelo GLM ..... 191

5.4.2.3. Diagnóstico del Modelo ..... 199

5.4.3. Principales Resultados Obtenidos bajo GLM ..... 201

**CAPITULO 6: CONCLUSIONES GENERALES ..... 207**

**BIBLIOGRAFIA ..... 215**

**INDICE DE FIGURAS ..... 226**

**INDICE DE TABLAS ..... 230**

**RESUMEN**

Análisis del Riesgo de Caída de Cartera en Seguros: Metodologías de "Inteligencia Artificial" vs "Modelos Lineales Generalizados"

Han transcurrido varios años desde que se comenzó a hablar de Solvencia II y hoy es una realidad; cuyo objetivo es el desarrollo y establecimiento de un nuevo sistema que permita determinar los recursos propios mínimos a requerir a cada aseguradora, en función de los riesgos asumidos y la gestión que se realice de ellos. Así mismo, engloba un conjunto de iniciativas para la revisión de la normativa existente, la valoración y supervisión de la situación financiera global de las entidades aseguradoras y modos de actuación interna de las mismas.

Uno de los temas más controvertidos bajo esta regulación es cómo conseguir una adecuada evaluación de los riesgos asumidos por las entidades. Esto se traduce en lograr identificar las causas que puedan suponer una pérdida en sus recursos; así como en innovar en el campo técnico para lograr una correcta cuantificación de los riesgos posibles en los que podrían estar expuestas las entidades.

El objetivo de este trabajo es mostrar la posibilidad de utilizar dos enfoques metodológicos distintos para la evaluación de riesgos: uno no paramétrico para lo cual se recurrirá a las técnicas de Inteligencia Artificial y, en contraste, la aplicación de los Modelos Lineales Generalizados provenientes de la estadística paramétrica. De esta forma, lograr establecer una serie de reglas de decisión básicas, a manera de herramienta de clasificación, que puedan ser capaces de determinar los perfiles de clientes susceptibles a la cancelación de su póliza.

La aplicación práctica de ambas metodologías, se llevará a cabo con la finalidad de analizar el Riesgo de Caída de Cartera; el cual hace referencia a uno de los tantos riesgos medibles que el sector habrá de tener en cuenta de acuerdo a Solvencia II. La

relevancia de ambas aplicaciones empíricas, será poder tener una aproximación a la probabilidad de cancelación del cliente mediante dichos patrones que se traduciría en una mejora en la gestión del Riesgo de Caída de Cartera; contribuyendo al equilibrio y estabilidad de los niveles de solvencia de las entidades.

La utilización de la Inteligencia Artificial es un enfoque novedoso para el análisis de dicho riesgo. Existen varios trabajos donde se han realizado diversas aplicaciones de la Inteligencia Artificial sobre infinidad de campos de estudio. Algunos se han enfocado en estudiar temas de gestión financiera o análisis de la solvencia de entidades. Sin embargo, existen muy pocas aplicaciones dentro del sector asegurador. Se han utilizado Máquinas de Vectores Soporte (SVM) para clasificar a los clientes del seguro de automóvil, atendiendo a si presentan o no siniestro en un período de un año (Tolmos, P. 2007). La teoría Rough Set se ha aplicado para estudiar la estabilidad financiera y predicción de la insolvencia dentro del sector (Sanchis et al., 2007; Shyng et al., 2007). Sin embargo, ninguno ha trabajado el tema del Riesgo de Caída de Cartera mediante estas técnicas.

De aquí lo novedoso del presente trabajo, ya que habitualmente se recurre a metodologías paramétricas del tipo estadístico y muy poco se ha indagado sobre la utilización de la parte no paramétrica que ofrece la Inteligencia Artificial.

Así, el trabajo se compone de seis apartados:

El **primer Capítulo**, sin ánimo de ser en absoluto teórico, se describe los fundamentos del proyecto Solvencia II, resumiendo sus antecedentes, así como los principales hitos y marco regulatorio. Se hace referencia a los tres pilares sobre los que descansa Solvencia II y que conforman la estructura de la normativa que, entre otros temas, busca gestionar la solvencia de las entidades bajo una visión homogénea y sin duda alguna está revolucionando al sector asegurador.

Bajo este contexto sobre el que surge la necesidad de estudiar nuevas metodologías de cuantificación y gestión de riesgos dentro del sector; la Caída de Cartera es un riesgo que, hoy en día, cobra especial importancia ante Solvencia II. Un **segundo capítulo** presenta sus principales aspectos, así como algunas de las

estadísticas del sector basadas en dicho riesgo, mencionando las principales causas que lo provocan.

En un **tercer Capítulo** se describe la información con la que se cuenta para las aplicaciones empíricas de las técnicas que se desarrollan más adelante, presentando el análisis exploratorio de las variables de la muestra, el tratamiento de los datos y el contexto de la información disponible.

Un **cuarto Capítulo** expone un modelo para lograr identificar patrones de comportamiento de clientes susceptibles a la cancelación de su contrato de seguros. Se revisan las metodologías no paramétricas propuestas por la Inteligencia Artificial, resumiendo las principales características de las técnicas de los Árboles de Decisión y Rough Set. Se realiza una aplicación empírica de ambas técnicas; lo que constituye una aportación metodológica como herramientas de predicción para este tipo de riesgo; finalizando con un resumen y análisis de los principales resultados obtenidos.

De igual forma, se comparan estos resultados en un **quinto Capítulo** dedicado a la aplicación empírica del modelo paramétrico. Se recopila las características, estructuras y componentes que engloban el marco teórico de los Modelos Lineales Generalizados. Posteriormente, se realiza su aplicación sobre la misma muestra de datos. Con base en los resultados obtenidos, se interpretan las conclusiones e implicaciones generadas a partir de este tipo de modelación predictiva, hoy también poco desarrollada en el problema en cuestión.

Finalmente, las **Conclusiones Generales**; donde sin ánimo de buscar el mejor ajuste y bonanza de los modelos aplicados; se ofrece la discusión de las principales conclusiones y resultados obtenidos, ofreciendo nuevas metodologías para ser utilizadas por las entidades aseguradoras.

Como se podrá observar a lo largo del trabajo realizado, existen muchos factores del propio negocio asegurador y visión comercial de cada entidad, que no han sido considerados y que seguramente deberían estar dentro de los supuestos que se hacen durante las aplicaciones empíricas realizadas.

Sin embargo, este trabajo no se centra en la fiabilidad y exactitud de los resultados obtenidos; sino la finalidad es animar a las aseguradoras a indagar en nuevas metodologías y técnicas, hasta hoy no del todo explotadas, para cubrir con las necesidades y requerimientos que Solvencia II le exigirá al sector.

**ABSTRACT**

Lapse Risk Analysis in Insurance: Methodologies of “Artificial Intelligence” vs “Generalized Linear Models”

Now, Solvency II is a reality after several years talking about it. Its objective is the development and establishment of a new system to ensure minimum capital requirements for each insurance company, depending on the risks assumed and the best management of them. It also includes a set of initiatives for existing legislation review; for evaluating and monitoring overall financial situation of insurers, as well as internal action procedures to control it.

With this background, one of the most controversial issues presented by Solvency II is how to get a proper assessment of risks assumed by the entities. First, entities have to identify the causes that may lead to a loss on entities resources, and then they have to innovate on technical fields for the best estimation of the potential risks they might be exposed.

The aim of this paper is to show the possibility of using two types of methods: a non-parametric using Artificial Intelligence techniques; in contrast to the results obtained with the parametric statistics by using Generalized Linear Models. Thus, to achieve a set of basic decision rules, as a classification tool that may be able to determine the profiles of policy customers susceptible to cancellation.

A practical application of both methodologies will be done in order to analyze Lapse Risk which refers to one of risk that insurance entities must take account under Solvency II regulatory. The relevance of both empirical applications will be able to have an approximation of probability of customer cancellation by those patterns. It would become an improvement in management of lapse risk, contributing to balance and stability of entity solvency levels.

Artificial Intelligence use is a novel approach to the analysis of risk. There are several studies where Artificial Intelligence applications have been done.

Some of them have been focused on financial management issues or entity solvency analysis. However, there are very few applications in insurance sector. Support Vector Machines (SVM) have been used to classify customer of automobile insurance, identifying if they have presented claim or not in a certain period of time (Tolmos, P. 2007). Rough Set theory has been applied to study financial stability and insolvency prediction in insurance sector (Sanchis et al., 2007; Shyng et al., 2007). However, no one has study Lapse Risk by these techniques.

Hence the novelty of this thesis, it´s used to use statistical parametric methodologies for it, and there has been done a very few research by using nonparametric techniques as Artificial Intelligence methods. Thus, the paper has been divided in six sections:

**First section**, no being theoretical at all, it will dedicated to general context that encourages the new methodologies study. Summarizing main backgrounds that lead the new Solvency II project. Regulatory framework that supports this scheme is presented. This new legislation is resumed by the 3 Pillars Structure that is reforming the insurance market.

Nowadays, Lapses risk is an important issue that faced the insurance market. Article describes a model in order to identify client behaviors that are susceptible to cancel its insurance policy. A **second chapter** presents the main aspects of Lapses risk considered according to Solvency II project; as well as some statistics based on this risk and the main causes that lead are also mentioned.

There is a **third chapter** where information for empirical applications is described. Thus, exploratory analysis of the variables is presented; also treatment given to the database and the context of information available is explained in this section.

In a **fourth chapter**, a model is exposed in order to identify patterns of behavior of customers susceptible to cancellation of your insurance contract. Thus, in this

section, the proposed Machine Learning methodology is reviewed, summarizing the main characteristics of Decision Tree Model and Rough Set techniques. After an empirical application is done and it's summarized the results obtained of this non-parametric technique.

Likewise, a **fifth chapter** is dedicated to empirical application of a parametric model. Therefore, after an introduction to this kind of methodology, a next section is dedicated to collect the main features, structures and components that resume theoretical framework of Generalized Linear Models. Then using the same data, it's proceeded to applicate the methodology offers by GLM. Based on results, it is presented some conclusions and implications that could be generated from this kind of predictive modelling.

Finally, a section of general **conclusions** of this paper are offered; where not with the exact fit of the models applied, it offers a new opportunity to investigate new methodologies to be used by insurers.

As it could be seen along the paper, there are many factors of insurance business and commercial vision of each insurance company that shouldn't have been considered during the empirical application of both methodologies.

However, the objective doesn't focus on the accuracy of results; but rather the aim is to encourage investigating new methodologies and techniques used by insurance industry. Many of them until today not fully exploited; which one are useful to cover requirements that Solvency II will require to the insurance sector.

## CAPITULO 1: PROYECTO SOLVENCIA II

### 1.1 Introducción

Desde que se comenzó a hablar en Europa de Solvencia II transcurrieron más de nueve años hasta que, a finales de 2009, se publicara la Directiva sobre Solvencia II. A partir de esta fecha, dicha Directiva ha seguido sufriendo diversas modificaciones, pero sus principios se han mantenido inalterables.

Al hablar de este proyecto no se debe olvidar mencionar a su análogo dentro del sector financiero, la banca, quien fue el pionero en buscar establecer un conjunto de recomendaciones y acuerdos sobre la legislación y regulación bancaria. Esto es, Basilea II, publicado en junio de 2004, y hoy superado por Basilea III publicado en diciembre de 2010; en los que se define la creación de un estándar internacional que sirva de referencia a los reguladores bancarios, con objeto de establecer los requerimientos de capital necesarios para asegurar la protección de las entidades frente a los riesgos financieros y operativos. Bajo una filosofía similar, se remontan los orígenes de Solvencia II, ya que sin dejar de tener en cuenta los matices propios del sector asegurador, la esencia y estructura de ambos proyectos es muy semejante.

Después de varios años y a pesar de sus prórrogas, Solvencia II es una realidad que ha puesto de manifiesto la adopción de medidas concretas para adaptarse al nuevo sistema; ya no es una práctica recomendable sino una materia exigible a las entidades.

De esta forma, Solvencia II, sin lugar a dudas, es un desafío para la industria aseguradora; quien tendrá que estar preparada para cuando sea plenamente aplicable, con carácter obligatorio, es decir, a partir del 1 de enero de 2016[1].

[1] *Directiva 2009/138/CE del Parlamento Europeo y del Consejo*, de 25 de noviembre de 2009, sobre el seguro de vida, el acceso a la actividad de seguro y de reaseguro y su ejercicio, en adelante *Directiva de Solvencia II*

Este proyecto es una verdadera transformación del modelo de gestión de riesgos y de la toma de decisiones en las entidades aseguradoras; por lo que el proceso de adaptación a esta nueva norma y filosofía exigirá un gran esfuerzo por parte de las entidades.

Existen varios motivos y necesidades que promueven el inicio de este proyecto; sin embargo el sector asegurador es consciente de que el objetivo esencial de Solvencia II era la "protección de los tomadores de seguros"; pero hoy por hoy han surgido nuevos objetivos tras la puesta en marcha de la Nueva Directiva de Solvencia II; como son: el mantenimiento de la confianza en el sistema financiero-asegurador, el fortalecimiento de la gestión de riesgos, la armonización de las prácticas de valoración contables y regulatorias y el común acuerdo sobre las medidas de intervención.

Todos estos conceptos e inquietudes han provocado que paulatinamente, el sector asegurador se cuestione sobre la innovación de los modelos y planteamientos en vigor; en otras palabras, se busca introducir al sector asegurador dentro de una nueva cultura del riesgo y su gestión; que se traduce en la Nueva Directiva de *Solvencia II*.

### 1.2 Antecedentes

Desde ya hace tiempo en algunos países surgió la inquietud de referenciar a la solidez financiera con los riesgos asumidos de manera implícita a la propia actividad de la entidad aseguradora.

En la década de los 50, los pioneros en la aplicación de esquemas basados en el riesgo fueron los finlandeses, quienes empezaron a utilizar un modelo de capital considerando el carácter estocástico de la actividad aseguradora mediante las "Reservas Especiales de Nivelación"[2]. Posteriormente, le secundó Canadá que a mediados de los 80 comienza a aplicar modelos que intentan englobar la totalidad de los riesgos mediante la generación de escenarios para el diseño de sus planes de negocio a través de las llamadas "Exigencias de Capital Mínimo para la Continuación"[3].

Bajo una línea similar, en los años 90, EE.UU. mediante la *NAIC*[4] desarrolla el modelo *RBC*[5] basado en un conjunto de normas haciendo una primera definición y basando los requerimientos de capital en una serie de riesgos independientes entre sí. Sin embargo, en este modelo, los activos y pasivos se valoran de acuerdo a las normas contables conocidas como *US-GAAP*[6], más no respecto al mercado; no se hace referencia a modelos internos o prueba de escenarios y no cuenta con algún elemento que sea equivalente a las exigencias contenidas en los Pilares II y III que propone Solvencia II.

---

[2] Término en inglés *"Special Equalization Reserves"* donde su constitución tiene como objetivo la estabilización de la solvencia frente a desviaciones de la siniestralidad a lo largo de los años

[3] Término en inglés *"Minimum Continuing Capital and Surplus Requirements"* que cita que el capital requerido debe ser determinado con base en 5 componentes de riesgos: asset default, life assumptions risk, interest rate changes, segregated funds and foreign Exchange risk

[4] *"National Association of Insurance Commissioners"* pone en marcha el sistema en 1993 para las entidades de Vida y en 1994 para las de No Vida

[5] *"Risk Based Capital"* que incluye los riesgos de inversión en renta fija, renta variable, riesgos de crédito y riesgos de suscripción (reserva de siniestros y reserva de siniestros pagados)

[6] "*US Generally Accepted Accounting Principles*" son los "Principios de Contabilidad Generalmente Aceptados" usados por las compañías con sede en Estados Unidos o cotizadas en Wall Street

De esta manera, se llega al año 2004 con el modelo suizo conocido como "Test Suizo de Solvencia"[7] que introduce la *FOPI*[8] comenzando un proceso de cambio en la supervisión aseguradora buscando un enfoque basado en el análisis de los riesgos reales que soporta una entidad aseguradora de una forma integrada.

Dicho esquema es muy semejante al de Solvencia II, ya que su esencia es muy parecida persiguiendo la protección del asegurado; tienen el mismo sistema basado en principios y se estructura en tres pilares buscando valorar tanto al activo como pasivo de acuerdo al mercado. De igual forma, se habla de tener un modelo estándar, aunque también se motiva a las entidades al uso de modelos propios.

Sin dejar de mencionar al modelo británico el cual funciona desde el año 2005 que también ha buscado relacionar los requerimientos de capital con los riesgos a los que están expuestos a las entidades. El sistema se basa en el cálculo de dos cifras: las "Exigencias Mejoradas de Capital"[9] y la "Evaluación del Capital Individual"[10]; las cuales deben ser comunicadas a la *FSA*[11] quien posteriormente decide el nivel de capital exigido para cada entidad aseguradora.

Así también se suele comparar la directriz de Solvencia II con la interpretación bancaria de un modelo similar como lo es Basilea III. Es cierto que comparten estructuras semejantes basándose en tres pilares enfocados en temas comunes; así mismo se habla de modelos internos para evaluar el riesgo o bien un modelo estándar. Sin embargo, la mayor de las diferencias es la forma de tratar los riesgos, ya que Solvencia II busca analizarlos de una forma integrada además de centrarse tanto en los riesgos expuestos del activo como en los del pasivo. Y fundamentalmente, es la naturaleza de la creación de ambos proyectos lo que los hace distintos. Por un lado, Basilea II busca fortalecer y estabilizar al sistema bancario; sin embargo, el marco de

[7] Término en inglés "*Swiss Solvency Test*" que comenzó desde mayo de 2003 pero no fue hasta 2004 cuando se elaboró el primer trabajo conceptual

[8] Oficina Federal Suiza de Seguros bajo el término en inglés *"Federal Office of Private Insurance"*

[9] Siglas en inglés "*ECR –Enhanced Capital Requirement"*

[10] Siglas en inglés "*ICA –Individual Capital Assessment"*

[11] Autoridad de Servicios Financieros bajo el término en inglés *"Financial Services Authority"*

Solvencia II está focalizado en la protección de los asegurados con quienes están comprometidas las entidades.

Se observa, por tanto, que el tema de la solvencia dentro del sector asegurador no es un tópico nuevo; así como su regulación; ya que existen varias directivas y conjunto de normas que se han ido perfeccionan y complementan entre sí. De ahí la necesidad de lograr establecer un conjunto de normas común que engloben la actual coyuntura, con el objetivo de adecuar la regulación a la situación actual, sin buscar cubrir una carencia sino completar las directrices ya existentes.

Los antecedentes más directos de Solvencia II se sitúan en su predecesor, Solvencia I. Estaba basado en un conjunto de ratios que relacionan el capital exigido con el volumen del negocio obtenido a partir del cálculo del Margen de Solvencia Obligatorio y el Fondo Mínimo de Garantía. Sólo se dirigía a los riesgos técnicos que surgen del pasivo de las entidades, sin tener en cuenta los riesgos asociados al activo como son las inversiones o la calidad crediticia de las operaciones. Sin dejar de mencionar que esta valoración y exigencias de capital no se hacen de acuerdo al mercado y sin considerar diversificación o transferencia de riesgos que implicase reducción de dichos requerimientos. Todas estas limitaciones dieron lugar a la necesidad de la creación de un nuevo modelo, materializado en la implementación de Solvencia II.

### 1.3 El Proyecto Solvencia II

Ante la necesidad de evolución del sector asegurador, éste se ve influenciado por la globalización y cambios normativos que están produciendo dentro del entorno financiero; que empiezan a condicionar y afectar indirectamente a las entidades aseguradoras.

Pues bien, es así como surge el interés en temas como la gestión de riesgos y solvencia con la que es capaz de enfrentarse una entidad aseguradora. Pero antes de seguir, es conveniente definir, primeramente, el concepto financiero del término solvencia. Se entiende por solvencia a la capacidad financiera de una entidad para hacer frente a todas sus obligaciones y cumplir con sus compromisos futuros (Alonso, Alberto A., 2008).

De esta manera, llevado este concepto al sector asegurador[12], se entiende como solvencia a la capacidad del asegurador para hacer frente a los compromisos y cobertura de riesgos adquiridos como consecuencia de la propia actividad aseguradora a través del volumen de sus reservas o provisiones matemáticas. En otras palabras, la solvencia de las entidades aseguradoras garantiza la indemnización de los asegurados en caso de pérdidas.

A partir de esta común interpretación sobre la solvencia de una entidad aseguradora, se puede referir que Solvencia II busca reunir un conjunto de normas relacionadas con la regulación de dicha solvencia de las compañías de seguros dentro del ámbito europeo.

Ahora bien, los análisis de solvencia que se venían realizando en la mayor parte de las aseguradoras, es decir el marco de Solvencia I, se basan en metodologías desarrolladas en los años 70[13]. La Directiva consideró que las entidades aseguradoras dispusieran de una reserva o capital complementario, además de las reservas técnicas, para hacer frente a los compromisos contraídos con sus asegurados. De aquí surge el

[12] *Ley 20/2015, de 14 de julio*, de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras

[13] En el año 1973, la Comunidad Europea dicta la Primera Directiva del Consejo *73/239/CEE*

término de *Margen de Solvencia*, el cual guardaría relación con el volumen global de la operaciones de la compañía y se calcularía en función de las primas y siniestros registrados por las entidades. Por otro lado, otro tema que aparece a partir de dicha Directiva es la necesidad de exigir un *Fondo Mínimo de Garantía*; esto es, un mínimo de seguridad por debajo del cual se pueda ver reducido el *Margen de Solvencia* y por tanto, la situación financiera de la compañía se encuentre en dificultades para cumplir sus compromisos.

Es así como estas ideas iniciales sobre el control de solvencia empiezan a adquirir importancia cuando a mediados de los años 90, se permite a las autoridades supervisoras considerar dichas medidas como alertas de incumplimientos por parte de las entidades aseguradoras que pudiesen perjudicar al asegurado.

A partir de aquí, es como se producen una serie de Directivas[14] posteriores que adoptan los diversos cambios con respecto a los requisitos de margen de solvencia que deben constituir las compañías de seguros. Con toda esta evolución, se estableció Solvencia I como un esquema sencillo que permite realizar un comparativo de los resultados financieros de las distintas entidades aseguradoras.

Sin embargo, estas normas no tomaban en cuenta la totalidad de los riesgos asumidos al que una entidad está expuesta; ya que la solvencia de una entidad se calculaba bajo la fácil aplicación de un conjunto de ratios basados en los niveles de siniestralidad y primas. Y por si fuera poco, estas normas no varían entre las distintos Estados miembros de la Unión Europea y son aplicadas de la misma forma por cualquier entidad aseguradora independientemente del tamaño y perfil de riesgo que asumen sus operaciones.

De ahí surgió la necesidad de un nuevo régimen que tenga en cuenta los avances más recientes en materia de supervisión y creación de nuevas técnicas actuariales para la medición y gestión de riesgos. El sector asegurador requiere un nuevo esquema regulatorio para evaluar la verdadera situación de solvencia de una entidad, buscando homogeneización y comparabilidad con el resto del mercado

[14] En el año2002, surge la Directiva 2002/13/CE del Parlamento y Consejo Europeo; entre otras.

financiero. Y todo ello, sin olvidar el objetivo principal, es decir, una mejor protección de los asegurados.

Es así como el sector asegurador se encamina hacia un sistema más complejo ante las necesidades y exigencias de su entorno. Empezando con todo un proceso de revisión; surge la propuesta por parte de la Comisión Europea de una nueva Directiva en materia de seguros y reaseguros, tanto del ramo de vida como de ramos distintos del de vida[15] bajo la denominación de "Solvencia II". Es a partir de aquí que este nuevo régimen busca establecer nuevos requisitos de solvencia a cumplir por parte de las entidades aseguradoras y revisa globalmente las condiciones financieras de dichas entidades, con el fin de obtener una mayor transparencia y convergencia dentro del sector asegurador.

Desde sus inicios, el proyecto Solvencia II se presenta como un gran reto que impone la Comisión Europea para lograr un ambicioso marco regulatorio de la supervisión de seguros en la Unión Europea. A modo de interpretación de los motivos que originan el planteamiento de un nuevo entorno normativo, Solvencia II surge de la confluencia de una serie de exigencias implícitas que empieza a necesitar el sector asegurador.

Como un primer objetivo, normativamente se aspira a una máxima armonización para conseguir un único mercado en el ámbito asegurador, delimitando detalladamente el contenido de las reglas que se adopten. Se busca mejorar el funcionamiento del sector asegurador mediante el establecimiento de normas coordinadas para la vigilancia de las entidades aseguradoras.

Sin embargo, esto no es un tema sencillo de abordar, ya que no es fácil conseguir un acuerdo sobre el qué y el cómo se va hacer y hasta donde se pretende llegar.

Seguido de ello, también se aspira a una actuación supervisora única en cuanto a prácticas y reglas de los distintos supervisores de seguros. Es decir, otro de los principales objetivos del régimen Solvencia II es eliminar las diferencias entre las

[15] Se excluyen las pequeñas mutuas y empresas de seguros. Así mismo, la Directiva no se aplica a los fondos de pensiones [COM(2008) 119]

legislaciones de los Estados miembros en lo relativos a las normas a que están sujetas las empresas de seguros y reaseguros

Así mismo, este nuevo entorno busca proteger a los acreedores mediante el establecimiento de procedimientos de saneamiento y liquidación de las entidades de seguros. En otras palabras, esto significa que para los negocios muy diversificados, equilibrados, actuarial y financieramente sólidos, Solvencia II debiera ser un buen negocio. Y lo contrario para los negocios excesivamente arriesgados o concentrados.

Para lograr conseguir estos objetivos propuestos por esta nueva normativa, es importante que el empresario de seguros conozca bien los riesgos inherentes a su actividad, sepa medir su magnitud probable en forma de carga de capital y, consecuentemente, estar adecuadamente provisto para hacer frente a ello. Es decir, financieramente hablando, su margen de solvencia debe ser un concepto dinámico, esto es, que derive cargas de capital distintas para distintos perfiles de riesgo.

En otras palabras, Solvencia II se engloba como un proyecto que asume la necesidad de una supervisión basada en todos los riesgos que una entidad aseguradora afronta en su negocio sobre la base de un patrimonio valorado de forma consistente con el mercado, junto con unos estándares máximos de calidad en la gestión de riesgos e información facilitada al mercado.

Sin dejar de mencionar que uno de sus mayores retos es la convergencia de dicha actividad supervisora; esto es, una de las labores esenciales es la de fomentar la aplicación coherente y unificada en la legislación y vigilancia de la actividad aseguradora. Por lo que el sistema consiste sólo en una serie de coeficientes e indicadores cuantitativos sino también debe considera el enfoque cualitativo que interviene en temas como el tipo de riesgos que asume la compañía, así como el tipo de gestión y control de los mismos. De esta forma, se proporciona a los supervisores herramientas apropiadas para evaluar la solvencia global de las entidades.

Bajo estos antecedentes y conociendo los orígenes de este macro proyecto, se puede empezar a suponer que los nuevos requerimientos de capital son mucho más altos de lo que originalmente se tenían previstos; ya que la eminente evolución del

proyecto ha hecho que sus principios hayan ido buscando robustez con el fin de lograr la fiabilidad que requiere y necesita el sector asegurador bajo este nuevo entorno normativo para mantener la competitividad que hasta ahora ha logrado sostener dentro del mercado financiero.

### 1.4 Marco Regulatorio

Como se ha mencionado, llegar a un acuerdo en la elaboración del conjunto de normas sobre el que descansará el entorno de Solvencia II, no ha sido un camino sencillo. El nuevo marco regulatorio debería ser tan eficiente y flexible como fuera posible; ante la continua evolución del sector asegurador en cuanto al desarrollo de nuevos productos, métodos y modelos.

Es por ello que se ha recurrido al denominado "Método Lamfalussy"[16], esto es, un modelo decisorio para la adopción y aplicación de los actos legislativos comunitarios en el sector de los servicios financieros: mercado de valores, bancos y seguros. De esta manera, se busca obtener de una forma rápida y eficaz la mayor convergencia posible en la legislación adoptada. Así su objetivo final es la aplicación coherente y completa de una normativa integrada en un mismo mercado comunitario.

El enfoque Lamfalussy estructura el proceso de puesta en marcha de las normativas europeas basándose en cuatro niveles cuyas características básicas se describen a continuación:

- **Nivel 1.** Tras un proceso completo de consulta, la Comisión Europea inicia el proceso de elaboración de una propuesta de Directiva (reglamento) que recoja los principios generales esenciales. Una vez que el Parlamento Europeo y el Consejo se ponen de acuerdo sobre estos principios de regulación del trabajo y sobre las competencias de ejecución de quienes se las ha encomendado la labor de preparar la norma, las medidas concretas de ejecución se desarrollan en el nivel 2.

- **Nivel 2.** Previa consulta técnica a su asesor de alto nivel *CEIOPS*[17], la Comisión Europea consulta con el comité de *EIOPA*[18] las medidas técnicas de ejecución. Así, el *CEIOPS* prepara su dictamen en consulta con los operadores del mercado, los usuarios finales y los consumidores, y se lo comunica a la Comisión

[16] Revisión del Proceso Lamfalussy en el comunicado *[COM(2007) 727 final - Diario Oficial C 55 de 28.2.2008]*
[17] Committee of European Insurance and Occupational Pensions Supervisors
[18] European Insurance and Occupational Pensions Authority

Europea. Ésta las examina y prepara una propuesta formal que somete al *EIOPA*, quien debe someterla a votación en un plazo máximo de tres meses. Si tal propuesta es aceptada por el *EIOPA*, la Comisión Europea adopta la medida. A lo largo de esta fase, se mantiene plenamente informado al Parlamento Europeo y se concede la máxima consideración a su opinión.

- **Nivel 3.** Corresponde a la tarea del *CEIOPS* de elaborar recomendaciones, normas y procesos comunes, interpretaciones conjuntas y directrices coherentes. También evalúa y compara la práctica reguladora para garantizar una implementación y aplicación coherente de manera que se logre una convergencia entre los métodos de supervisión.

- **Nivel 4.** La Comisión verifica el cumplimiento de la normativa comunitaria por los Estados miembros y puede emprender actuaciones judiciales contra los Estados miembros que presuntamente infrinjan el Derecho Comunitario.

Bajo este planteamiento se pondría obtener una rápida toma de decisiones conjunta con una consulta completa y transparente de todos los miembros interesados.

De esta forma, se incluyó el desarrollo de dicho *enfoque Lamafalussy*; resumiendo un conjunto de 42 medidas destinadas a provocar cambios sustanciales en la regulación de los mercados financieros de la UE y completar un mercado único. Como resumen, se puede decir que los objetivos específicos del *FSAP* son:

- Lograr un mercado único a escala mayorista
- Lograr un mercado abierto y seguros para los consumidores y
- Disponer de normas reguladoras y de supervisión actualizadas.

En otras palabras se puede decir que, desde los inicios del planteamiento del marco regulatorio sobre el que descansa el proyecto Solvencia II, se busca que las actuales Directivas den soporte a una legislación para los Estados miembros y permitan que las entidades aseguradoras cuenten con un mismo marco para poder prestar sus servicios en otros países de la UE.

Sin embargo, existen múltiples áreas en las cuales la legislación pudiera dar pie a múltiples interpretaciones debido a la ausencia de criterios y normas claras y precisas; y por tanto, verse alejado del objetivo inicial, la creación de un mercado único bajo criterios armonizados y normas comunes relativas a la solvencia para toda la UE. He aquí la actual situación en que se encuentra el nuevo entorno normativo.

Las autoridades supervisoras como *EIOPA* intentando mitigar dicha incertidumbre en la gestión del sector asegurador, proponiendo una serie de normas transitorias que aprovechan lo hasta ahora invertido y desarrollado por el sector.

A partir de aquí surgen las siguientes medidas que deberán ser previstas y aplicadas por todas las entidades aseguradoras bajo el marco normativo de Solvencia II:

- **Gobierno Corporativo**

Las entidades deben formar un sistema de gobierno documentado que busque la eficacia mediante la definición de un proceso de actuación para la gestión del negocio de manera prudente y en línea con la magnitud y complejidad del diseño de riesgos a los que la entidad está dispuesta o pretende asumir.

- **Autoevaluación de los Riesgos (proceso ORSA[19])**

Para este proceso se establecen una serie de principios definidos como:

i) Involucración del Consejo de Administración y de la Alta dirección durante el proceso

ii) Todos participan en el proceso: Áreas claves, financieras, estratégicas y de negocio conjuntamente con el Consejo de Administración

iii) El proceso ORSA no es solo un informe, su importancia radica en el propio proceso de creación del mismo

[19]Según sus siglas en ingles, *Own Risk and Solvency Assessment*

iv) Integrar y conjugar el proceso ORSA como parte del negocio

v) Gestionar el día a día de la entidad con base en lo definido en el proceso ORSA

vi) Planificar y contemplar la realización de varias versiones del proceso antes ser el definitivo

- **Informe al Supervisor**

Se estima presentar el primer informe al supervisor con datos a finales de 2015 utilizando un conjunto de plantillas predefinidas. Todo ello, con la colaboración de los Supervisores locales para la asistencia y soporte en el período de transición hacia Solvencia II, con el fin de minimizar posibles errores en la información solicitada.

- **Pre-aplicación de modelos internos**

Se pide a los Supervisores locales que conozca el grado de preparación de las entidades en los Modelos Internos; así como facilitar la emisión de cierta retroalimentación sobre los avances o estado de situación de las compañías aseguradoras.

### 1.5 Los 3 Pilares que la componen

Solvencia II es la gran apuesta de la Comisión Europea, y de su asesor de alto nivel el CEIOPS, hoy convertido en autoridad supervisora europea con las siglas EIOPA.

Al igual que lo hiciese su análogo, Basilea II, este nuevo acuerdo normativo, Solvencia II, se compone de un conjunto de elementos que se ordena bajo una estructura basada en tres pilares (*Figura 1*), que se resumen de la siguiente forma:

- Pilar I – *Cuantitativo*: Se destina a los requerimientos cuantitativos. El objetivo es determinar el "Balance Económico" enfocado al Riesgo propio de la entidad y valorado a Mercado mediante normas establecidas para la valoración de los activos y pasivos con los que cuentan las entidades aseguradoras.


**Figura 1:** *Esquema conceptual del Proyecto Solvencia II*
**Fuente**: *Propia de los autores*

- Pilar II – *Cualitativo*: Se destina a los requerimientos cualitativos y las normas de supervisión. Busca una supervisión de alta calidad por parte de los organismos reguladores, con rigurosas exigencias en materia del gobierno en las entidades aseguradoras, que afectan a los órganos de gestión y dirección de la misma quienes son los principales responsables de los procesos de identificación, medición y gestión activa del riesgo. De esta forma, se ven obligadas a buscar mejoras en la gestión interna y así conseguir reforzar la estabilidad y solvencia del sector asegurador.

- Pilar III – *Disciplina del mercado*: Se busca desarrollar la comunicación de la información entre el supervisor y la entidad aseguradora, con el fin de favorecer la disciplina, transparencia y así lograr conseguir una mayor estabilidad financiera mediante una tendencia hacia la obtención de una contabilidad internacional homogénea.

### 1.5.1 Pilar I – Cuantitativo

En este apartado se analiza el contenido del Pilar I y las principales implicaciones asociadas al cumplimiento de los requisitos exigidos para las entidades aseguradoras y definidas en el mismo.

Este primer Pilar I se le ubica como el pilar técnico de Solvencia II; ya que va desde formulaciones relativamente sencillas hasta otras de una enorme complejidad. En esta fase se busca la construcción de un primer Pilar matemáticamente sólido y, a la vez, integral, en el sentido de ser capaz de abarcar la totalidad de riesgos que se deben calcular en una adecuada valoración del perfil de riesgo de un negocio.

Las entidades aseguradoras deben ceñirse a una serie de requisitos exigidos dentro de las normas establecidas en el Pilar I; las cuales se pueden englobar en los siguientes 6 principales materias de actuación:

- Valoración de activos y pasivos: Se refiere a las normas establecidas que indica de qué manera debe calcularse el balance económico de las entidades. Se hace

referencia al fortalecimiento del actual tratamiento del activo y pasivo basando su valoración a valor de mercado.

- Cálculo de las provisiones técnicas: Se busca establecer provisiones técnicas destinadas a garantizar que las entidades cumplan con sus obligaciones frente a sus asegurados. Se recurre al desarrollo de metodologías actuariales sobre el cálculo de dichas reservas con base en el método del "Mejor Estimador".

- Valoración de los fondos propios: Se refiere a los recursos financieros que las entidades poseen para hacer frente a los riesgos y absorber las pérdidas posibles. Definir un esquema de cobertura basado en la calidad de sus recursos propios.

- Cálculo del capital de solvencia obligatorio: Se apunta sobre los fondos que las entidades necesitan para limitar la probabilidad de ruina, el cual estará sujeto a un control continuo. Establecer un esquema que recoja la totalidad de los riesgos a los que está expuestos en concordancia con su perfil de riesgo específico.

- Cálculo del capital mínimo obligatorio: Se hace mención sobre los fondos propios de base admisibles para cubrir el nivel mínimo por debajo del cual los intereses de los asegurados podrían verse afectados y necesario para que las entidades puedan seguir desarrollando su actividad.

- Definición de las reglas de inversión: Se refiere al control, gestión e inversión de los activos en poder de las entidades, las cuales deben llevar a cabo resguardando los intereses de sus asegurados. Instaurar un esquema de inversión congruente con la naturaleza de sus pasivos favoreciendo a la conservación de un nivel de activos vs pasivos adecuado.

Para cumplir con estos requisitos, es necesario que primeramente, las entidades aseguradoras establezcan y definan las reglas de valoración que seguirán para cuantificar todas las partidas relevantes del balance económico., tanto del activo como del pasivo. Dicha valoración se deberá realizar en conjunto de tal forma que se pueda obtener los niveles de capital adecuados al perfil de riesgos asumidos por la entidad.

Es así como surge un término a tener en cuenta a partir de este punto: Capital de Solvencia Obligatorio *(SCR – Solvency Capital Requirement).* Se define como el capital necesario para hacer frente a las posibles pérdidas económicas teniendo en cuanto todos los riegos cuantificables a los que está expuesta, en un horizonte temporal de un año y con un nivel de confianza del 99.5% (*VaR al 99.5%*)[20].

Para la cuantificación de dicho Capital Requerido, Solvencia II facilita su propia metodología denominado *"modelo estándar*"; o bien permite a cada entidad implementar un *"modelo interno*" basado en la experiencia propia de la compañía.

En términos generales, el "*modelo estándar*" establece una fórmula general para el cálculo del SCR; el cual fue definido por el Comité Europeo de Supervisores de Seguros y de Pensiones de Jubilación (*CEIOPS*[21]). De esta forma, la valoración del requerimiento de Capital se obtiene mediante el desglose de seis sub-módulos de cálculo correspondiente a la valoración de los riesgos asumidos por la entidad.

Por el contrario, los "*modelos internos*" deberán construir sus propias hipótesis basadas en la experiencia de la compañía, justificando y documentando cada una de éstas, así como la estructura y calibración de cada riesgo considerado. Así mismo, si la entidad opta por utilizar su propio modelo, éste deberá ser presentado y aprobado por los órganos supervisores.

#### 1.5.1.1 Fórmula Estándar

Se resumen con un conjunto de normas que asume un enfoque general de identificación y valoración de los riesgos que afectan a las entidades aseguradoras. Con base en dicha evaluación, se cuantifican cada uno de los riesgos y se calcula el capital necesario para cubrirlos.

Dentro del planteamiento que sugiere la Fórmula Estándar, se identifican los riesgos más relevantes de la entidad aseguradora, tanto en la estimación de las

[20] *Ley 20/2015, de 14 de julio*, de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras
[21] En la actualidad EIOPA, *European Insurance and Occupational Pensions Authority*

provisiones técnicas como en la valoración de los activos bajos la generación de diferentes escenarios.

En lo que se refiere a la parte de los activos, se valoran a valor de mercado teniendo en cuenta cualquier diferencial entre el precio en que se compra o el que se vende el activo. En los casos en que no se disponga del precio, se puede utilizar cualquier otro mecanismo de valoración que sea consistente con su valor en el mercado financiero.

Por la parte de los pasivos, se recurre a la utilización del método del "Mejor Estimador" Se busca obtener el valor más probable que tendrán las reservas técnicas de acuerdo a los escenarios futuros previamente definidos. Para ello, se obtiene el Valor Presente de los Flujos Futuros de la cartera asegurada que se estima obtener bajo las Hipótesis Actuariales basadas en la experiencia actual de la entidad. Algunas de dichas hipótesis son específicas de cada ramo de seguro y otras variables pueden ser definidas de manera general.

Para lograr el objetivo del enfoque que ofrece la Fórmula Estándar, fue necesario dejar definidos los criterios de valoración y la clasificación de los riesgos por ramos. Pero además de ello, ha sido necesario, establecer una medida de riesgo con la cual se calculan los requerimientos de capital bajo los distintos escenarios definidos. Es decir, el importe necesario para alcanzar un nivel de confianza adecuado para cubrir posibles contingencias ante las cuales se pueda ver amenazado el patrimonio de la entidad[22].

De esta forma, el cálculo del requerimiento de capital se puede obtiene mediante el desglose de seis sub-módulos de cálculo correspondiente a la valoración de los riesgos asumidos por la entidad. Y posteriormente mediante una matriz de varianza y covarianzas, realizar la agregación de riesgos y obtener el *SCR* (*Solvency Capital Requirement*) global, es decir, el Requerimiento de Capital de Solvencia bajo el enfoque de la Fórmula Estándar (*Figura 2*).

[22] *Ley 20/2015, de 14 de julio*, de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras

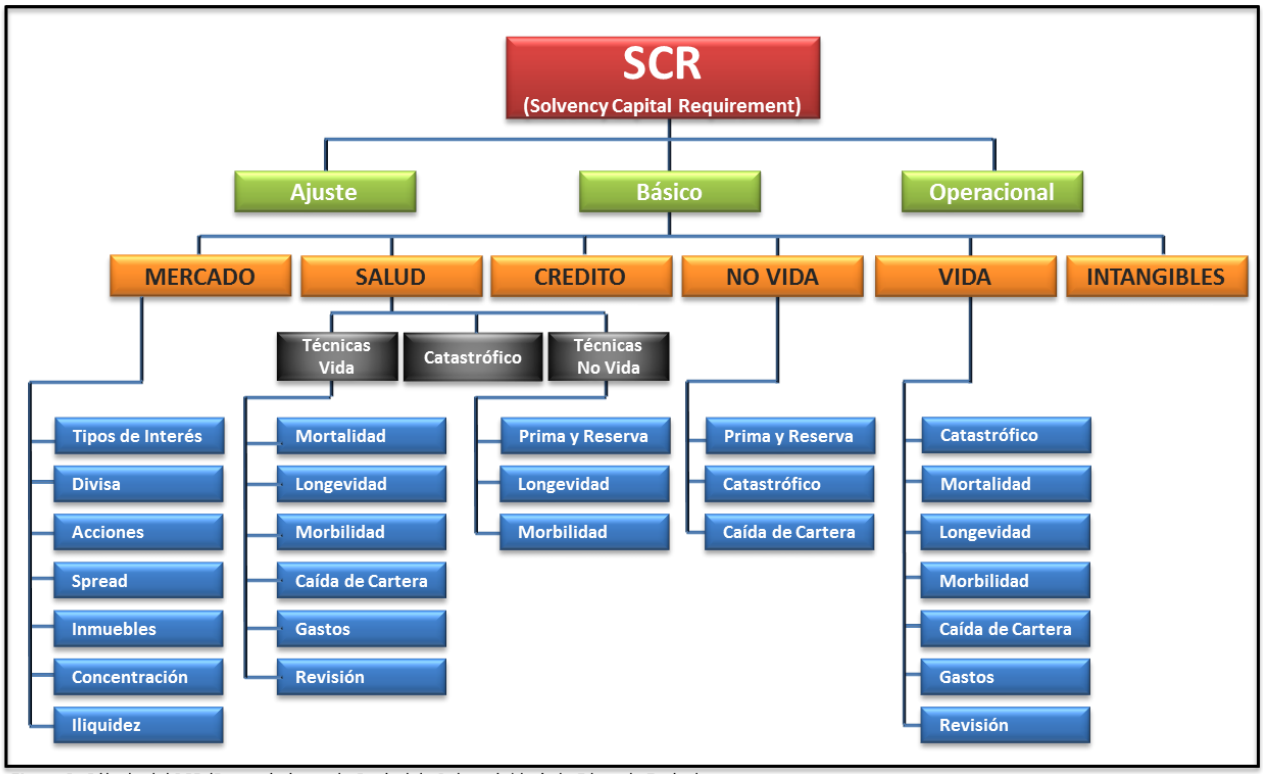


**Figura 2:** *Cálculo del SCR (Requerimiento de Capital de Solvencia) bajo la Fórmula Estándar*
**Fuente**: *Propia de los autores*

### 1.5.1.2 Modelos Internos

Solvencia II establece un nuevo enfoque para determinar los niveles de solvencia de las entidades aseguradoras. Es decir, los requisitos de capital debe reflejar la capacidad de las compañías de seguros de afrontar sus obligaciones durante un intervalo de tiempo definido y bajo un nivel de confianza establecido; tomando en cuenta todos los riesgos a los que está expuesto: técnico, operativo, inversión, mercado, crédito, etc.)

A partir de la definición de dicho enfoque de cálculo surge el planteamiento de emplear el método estándar, previamente expuesto; o bien proponer el diseño y utilización de un modelo interno propio al perfil de riesgo asumido por cada entidad aseguradora. De esta manera, se transfiere la competencia a las compañías aseguradoras de calcular su capital de solvencia en base a los verdaderos niveles de riesgo que le apetece y es capaz de asumir.

El punto de partida de cualquier modelo interno debe ser la generación de escenarios de riesgos[23] dentro de un horizonte temporal, generalmente referido a un año vista. Para ello es necesaria la medición de la solvencia en cada escenario mediante el valor de los activos y pasivos dentro de este año; y sus posibles cambios de valor a lo largo del tiempo. Y luego bien, a partir de aquí, establecer la medida de riesgo y nivel de confianza que se utilizará para finalmente obtener los requisitos de capital exigidos.

El nuevo entorno de Solvencia II busca fomentar el desarrollo de modelos internos de riesgo para el cálculo de su nivel de capital de solvencia requerido. La normativa presenta un marco flexible en cuanto a la elección de modelos internos. Esto es, admite el uso de modelos internos, tanto totales como parciales; es decir si mezcla elementos de la Fórmula Estándar en cualquier de los módulos del mapa de riesgos, tanto por un lado del activo como por lado del pasivo.

Es así como la Directiva marca ciertos requisitos para la implementación de los modelos internos. Por un lado, la entidad no sólo debe demostrar que el modelo interno es ampliamente utilizado; sino que también desempeña un importante papel en su sistema de gobierno; es decir, en su sistema de gestión de riesgos y toma de decisiones, procesos de evaluación y asignación del capital económico. Para ello, la directiva establece que los métodos usados para la determinación de la distribución de probabilidad estarán basados en técnicas actuariales y estadísticas adecuadas y coherentes; con información actual y fiable que serán utilizados bajo supuestos realistas.

Ahora bien, no se habla de métodos concretos para la determinación del capital económico; sin embargo el modelo interno debe clasificar el riesgo adecuadamente para garantizar que contemple todos y cada uno de los riesgos a los que la entidad está expuesta y, como mínimo debe considerar los riesgos considerados en el modelo estándar para el caso de los modelos internos completos.

[23] Coloquialmente referidos como “Escenarios Real World” que resumen diversos contextos macroeconómicos bajo diversos riesgos (financieros, biométricos, medioambientales, operativos, etc.) a los que se encuentra expuesta cada entidad aseguradora

Otro requisito que los modelos internos deben tener en cuenta es la adopción de medidas de gestión futuras que se prevén ante ciertos escenarios futuros indicando el tiempo de ejecución de dichas medidas. Se podrá tener en cuenta técnicas de mitigación de riesgos, siempre y cuando esto se vea reflejado adecuadamente dentro de los riesgos derivados de la mitigación, por ejemplo dentro del riesgo crediticio que esto supone.

Así también se habla como requisitos de definir un horizonte temporal y utilizar el *VaR* de los fondos propios al 99.5% para calcular el SCR. La entidad deberá comprobar el funcionamiento del modelo interno a través de cierto período, es decir, una especie de validación del modelo verificando que las hipótesis y especificaciones técnicas continúen siendo las adecuadas y comparando los resultados obtenidos por el modelo vs la realidad observada.

Estos son, en términos generales, algunos de los requisitos de los modelos internos, sin embargo, más que una receta secreta, lo importante es la esencia de lo que debe aportar los modelos internos a las entidades. Los modelos internos pueden contribuir a que la entidad desarrolle su actividad de forma más eficiente, identificando las áreas de negocio más rentables y facilitando la adecuada mitigación de riesgos.

### 1.5.2 Pilar II – Cualitativo

El régimen propuesto por Solvencia II exige que se lleve a cabo una supervisión, a fin de garantizar ante todo la protección de los asegurados. Por otro lado, teniendo en cuenta la estabilidad financiera y la equidad de los mercados, las autoridades de supervisión deben evaluar la situación financiera, así como los procesos realizados y metodologías adoptadas por las entidades para la gestión de sus riesgos. Para ello, los supervisores deben ejercer sus facultades en el momento oportuno y respetando el principio de proporcionalidad; es decir, evitar la utilización desmedida de las normas, haciendo uso exclusivo de éstas para protección y garantía de los asegurados.

Por tanto, el contenido de las exigencias expuestas dentro de este segundo Pilar sobre el que descansa la filosofía propuesta por Solvencia II muestra especial interés ante el incumplimiento de los requisitos cualitativos: gestión de riesgos y proceso supervisión adecuado. Es decir, busca inducir a las entidades aseguradoras a seguir principios sólidos sobre el control interno y resume un conjunto de recomendaciones con el objetivo de mantener una administración de los riesgos adecuada dentro de cada entidad aseguradora.

Por un lado, se centra en la implementación de un proceso de supervisión cuyo objetivo es el garantizar y evaluar una apropiada gestión empresarial. Se hace mención al establecimiento de una serie de criterios que sirvan de indicadores preventivos. Es decir, se refiere a fomentar una supervisión prudencial destinada a detectar aquellas entidades que presentan un riesgo elevado, por sus características financieras, organizativas o de cualquier otra índole; ya que ello podría tener graves consecuencias sobre la solidez financiera de las entidades.

Por otro lado, el Pilar II de este proyecto, hace especial hincapié en la necesidad de preservar la coherencia entre las exigencias impuestas entre los distintos elementos que conforman el sector financiero; como son la gestión del riesgo, solvencia, auditorías y controles internos dentro de cada entidad. Todo ello, en busca de un proceso de inspección por parte de las autoridades supervisoras que contemple:

- Coordinación de la acción de supervisión en épocas de crisis
- Competencias y medidas claramente definidas en momentos de intervención por parte de las autoridades supervisoras
- Transparencia y responsabilidad definida de la acción supervisora

En otras palabras, no es otra cosa más que la responsabilidad del cumplimiento de los requisitos cualitativos y de control que recae sobre los órganos de administración o dirección de las entidades aseguradoras; es decir el gobierno corporativo sobre el que descansa la entidad aseguradora.

### 1.5.3 Pilar III – Disciplina del Mercado

En lo que se refiere al contenido del Pilar III, se ha ce referencia a la obligación de las entidades a comunicar cualquier información a las autoridades de supervisión; de esta forma, el marco normativo propuesto por Solvencia II reúne un conjunto de principios que buscan ajustar y definir los lineamientos de la entrega de información sobre las entidades destinada al público.

Mediante la implementación de este tercer Pilar, se verá reforzada la transparencia de la actividad aseguradora, así como la solidez de la supervisión del seguro que se traduzca en el fortalecimiento de la disciplina del mercado financiero.

Ahora bien, los requisitos de información que contemplará este pilar dependen en gran medida de la descripción definitiva de las medidas adoptadas en el primer y segundo Pilar. Es por ello, que la fase de discusión y coordinación de las exigencias de información del proyecto de Solvencia II tomó bastante tiempo, mismo que propició que su fecha de implantación fuese postergada en numerables ocasiones.

Aunque la fecha de implantación de Solvencia II sufrió varias modificaciones y retrasos, los requisitos establecidos dentro del Pilar III se exigirán de forma progresiva; ya que existirá un período transitorio para la remisión de la información requerida dentro de este apartado. Todos los años, previa aprobación de la dirección administrativa, las entidades deberán publicar un informe en que presenten su situación financiera y de solvencia. Las entidades deberán aportar información actualizada y, si lo desea, toda la información adicional que considere oportuna y de interés de cara al supervisor y al mercado.

Uno de los grandes retos a los que se enfrenta las exigencias del Pilar III, es el fomentar la convergencia y transparencia de la actividad supervisora; ya que esto supone que se apliquen un conjuntos de normas establecidas dentro una legislación

única y comunitaria para todos los Estados miembros[24]. Es decir, es evidente que la forma en que las entidades aseguradoras se someten a supervisión es un factor clave para el éxito del mercado único y del régimen Solvencia II. Es por ello, que otra de las propuestas de este nuevo entorno normativo es el de introducir el concepto de "supervisor de grupo". Esto es, para cada grupo, se designará a una autoridad única a la que se conferirán facultades concretas de decisión y coordinación. Dichas facultades como son: la solvencia de grupo, concentración de riesgo, etc.; se ejercerán en cooperación con las autoridades de supervisión locales.

[24] Cabe mencionar la importancia que tiene la figura del Comité Europeo de Supervisores de Seguros y Pensiones de Jubilación (CESSPJ) que promueve una aplicación coherente de la propuesta Solvencia II y la convergencia de las prácticas supervisoras en Europa

## CAPITULO 2: RIESGO DE CAÍDA DE CARTERA

### 2.1. Introducción

Empezando con el proceso de revisión en el año 2001 y su primer fase concluida en al año 2003; es como surge la propuesta por parte de la Comisión Europea de una nueva Directiva en materia de seguros y reaseguros, tanto del ramo de vida como de ramos distintos del de vida[25] bajo la denominación de "Solvencia II". Es a partir de aquí que este nuevo régimen busca establecer nuevos requisitos de solvencia a cumplir por parte de las entidades aseguradoras y revisa globalmente las condiciones financieras de dichas entidades, con el fin de obtener una mayor transparencia y convergencia dentro del sector asegurador.

Solvencia II pretende que las entidades aseguradoras mantengan un volumen total de provisiones técnicas y un capital de solvencia que garantice su estabilidad ante fluctuaciones externas adversas. En definitiva, intenta que las compañías mantengan un nivel económico acorde con los compromisos asumidos, y que garantice la protección del asegurado (Ferri, et al 2010).

Ante esta nueva regulación, las compañías aseguradoras están siendo sometidas a desarrollar nuevas técnicas para la cuantificación y control de los riesgos a los que se encuentran expuestas. Todo ello con el fin de lograr implementar una gestión integral del riesgo que contemple un adecuado nivel de solvencia.

Dicha gestión de riesgos implica contemplar todos y cada uno de los componentes del negocio asegurador que puedan generar algún tipo de contingencia para la compañía. De esta forma, este nuevo proyecto regulador, logra identificar una serie de riesgos a los que podría estar expuesto el sector asegurador en cierto intervalo de tiempo. Uno de dichos riesgos contemplados es la caída de cartera que registra una entidad entendiéndose como tal a la rotación o salida de asegurados, lo

[25] Se excluyen las pequeñas mutuas y empresas de seguros. Así mismo, la Directiva no se aplica a los fondos de pensiones [COM(2008) 119]

cual se ve directamente reflejado en el decrecimiento en el volumen de primas de la entidad.

Bajo otra perspectiva, en un competitivo mercado asegurador donde cada día toman relevancia temas como la guerra de precios, accesibilidad a múltiples cotizaciones, así como la constante innovación en el desarrollo de productos; surge la necesidad de retener y fidelizar a los clientes. De esta forma, ya no sólo se presta atención a niveles de primas altos sino a la capacidad de garantizar la rentabilidad de la entidad, lo cual no es una tarea fácil ante una situación de crisis financiera como la que se vive actualmente.

Así pues, se sabe que todo tipo de variación, tanto en el volumen como en los ratios de rentabilidad, que pueda sufrir una entidad por este tipo de eventos se traduce en un riesgo considerable. La cuantificación de dicho riesgo de negocio es un tema fundamental dentro de la administración de riesgos de una compañía aseguradora. Así mismo, el control y mejora de dicho cálculo supone obtener niveles de requerimientos de solvencia óptimos para la compañía.

De aquí la importancia de la cuantificación del riesgo de caída de cartera que exige la nueva regulación de Solvencia II; así como los principales agentes causantes de su constitución e implicaciones que se reflejan directamente sobre los márgenes de solvencia de la entidad. Estos temas, así como la notación para el cálculo de los porcentajes de caída de cartera, serán algunos de los puntos a tratar en este capítulo.

## 2.2. Riesgo de Caída de Cartera

Para encuadrar el Riesgo de Caída de Cartera, se debe hacer referencia al Pilar I-Cuantitativo; el cual en otras palabras, busca establecer un sistema de gestión integral de riesgos mediante una mejora en el control y cuantificación de los mismos a los que están expuestas las entidades aseguradoras; que a su vez, se verá reflejado en términos de Capital Requerido para hacer frente a las obligaciones asumidas.

Recordando, las entidades podrán recurrir a la utilización del enfoque que propone la *Fórmula Estándar*; para la cuantificación de dicho Capital Requerido; es decir, el *SCR* (*Solvency Capital Requirement*). Bajo este enfoque, algunos autores engloban el cálculo del módulo del *SCR* como la agregación de riesgos resultado de la suma de: 1) Riesgos Suscripción: subdividido por ramo (Vida, No Vida, Salud), 2) Riesgo de Mercado y 3) Riesgo de Crédito (Ayuso et al 2012).

Siendo así, a efectos del presente trabajo, se debe focalizar en el "Riesgo de Suscripción" en los Seguros de Vida que contempla los riesgos técnico-actuariales asumidos por la entidad ante cualquier desviación de los parámetros biométricos considerados.

Por otro lado, los modelos internos establecidos en Solvencia II no se basan en ninguna fórmula específica para la cuantificación de los riesgos, sino que esta valoración deberá ser obtenida bajo hipótesis propias y metodologías adecuadas al perfil de riesgo de cada compañía.

Ahora bien, para la cuantificación del Riesgo de Caída de Cartera al que está expuesta una compañía aseguradora que opera el ramo de Vida; se habla de pólizas de Seguro de Vida, las cuales pueden ser temporales; cuando su vencimiento está previamente definido dentro del contrato de seguro; o bien, pueden ser renovables; en el caso de las pólizas que se renuevan cada cierto período por voluntad del asegurado.

Pues bien, se dice que la decisión de dicha renovación está directamente influenciada por la satisfacción del cliente; es decir la confianza y otros elementos

subjetivos cobran vital importancia frente a componentes más objetivos como son el coste del seguro o niveles de suma asegurada.

Siendo así que surge el concepto de "Caída de Cartera", del cual no existe una definición precisa; por lo que puede precisar como el conjunto de pólizas que no optan por la renovación a su vencimiento por parte de los asegurados (Millán y Colomina, 2001). Esto a su vez, se traduce en una fluctuación del volumen de negocio y los márgenes de solvencia. Es así como surge la necesidad de estudiar dicho evento como parte de los riesgos al que puede enfrentarse una entidad y por tanto, la importancia de la cuantificación y control del mismo.

En términos matemáticos, el Número de Pólizas que se anulan o cancelan durante un período determinado, se puede expresar de la siguiente manera:

$$Anul = Pol_i + Pol_{NP} - Pol_f$$

siendo: $Pol_i$= N° Pólizas en Vigor al inicio del período

$Pol_{NP}$= N° Pólizas de Nueva Producción registradas durante el período

$Pol_f$= N° Pólizas en Vigor al final del período

De esta forma, se puede expresar el concepto de Caída de Cartera en términos de porcentajes de la siguiente manera:

$$TasaCaída = \frac{Anul}{Pol_i}$$

Cabe mencionar que la cuantificación de la Caída de Cartera se puede realizar con base el N° Pólizas; o bien, es interesante analizar el impacto de las anulaciones por nivel de Prima o Reservas. De esta manera, el cálculo arroja un resultado en visión económica complementaria y mucho más robusta a la que ofrece el sólo análisis del volumen de pólizas que salen de la entidad.

En la actualidad, cada entidad aseguradora ha ido desarrollando metodologías novedosas con el objetivo de estimar la Caída de Cartera que se registrará en un futuro. En la mayoría de los casos, basándose en su información histórica, utilizan

modelos estadísticos medianamente complejos, y determinan los porcentajes de caída que definan mejor el riesgo al que está expuesto. Por otro lado, algunos autores han decidido utilizar el promedio de dichos porcentajes de caída obtenidos en los distintos períodos en la elaboración de escenarios de Caída; proponiendo escenarios extremos de dicho riesgo teniendo en cuenta el grado de contagio de las cancelaciones (Ayuso et al., 2011).

Cualquiera que sea la metodología utilizada, Solvencia II establece que las compañías aseguradoras deben estar cubiertas de los riesgos a los que está expuesta mediante la determinación del capital de solvencia obligatorio ante cualquier escenario extremo, obtenida a partir de la medida del Valor en Riesgo (*VaR*) de cada riesgo. Dicho valor en riesgo deberá tener en cuenta las desviaciones producidas por la ocurrencia de cierto escenario adverso, con respecto al "*Mejor Estimador*" (*Best Estimate*) que tenga la entidad sobre la frecuencia y severidad del riesgo.

Es aquí, donde se introduce el término de "*Mejor Estimador*" de acuerdo con la nueva regulación. Éste corresponde al Valor Presente Esperado de los Flujos de Efectivo Futuros utilizando supuestos técnico-actuariales que "mejor estimen" el comportamiento futuro de los riesgos biométricos (mortalidad, longevidad, caída de cartera, etc.); que pueda impactar a la cartera en vigor de la entidad; descontados a una tasa de interés libre de riesgo. En términos económicos, se traduce en la provisión que la entidad deberá tener en su pasivo para hacer frente a las obligaciones futuras que derivan de la suscripción de dichos riesgos.

Por un lado, se sabe que las probabilidades de cancelación varían dependiendo del ramo (Guillén et al. 2008). O bien, de acuerdo a los resultados obtenidos por algunos autores, se ha podido concluir que es totalmente válido realizar segmentaciones por antigüedad del cliente o tipo de producto (Ayuso et al. 2011).

Todo ello sugiere que el cálculo del mejor estimador puede venir determinado por diversos factores. De aquí, que surge el interés de este estudio, que plantea el análisis de la caída de cartera desde el punto de vista de los posibles factores que la promueven. Es decir, profundizar en el estudio del comportamiento del asegurado y ser utilizado como herramienta de gestión del riesgo de Caída de Cartera.

### 2.3. Estadísticas y Causas del Riesgo de Caída de Cartera

Como ya se ha puesto de manifiesto, la Caída de Cartera es un tema que cobra relevancia dentro del sector asegurador ante el proyecto Solvencia II. Es por ello, que resulta interesante revisar primeramente una serie de cifras e índices históricos que demuestran el comportamiento de este riesgo en el los últimos años.

Con el fin de exponer el estado en cuestión del riesgo de Caída de Cartera, es importante recoger cierta información del estudio realizado por ICEA[26] en el año 2013 sobre la caída de cartera en Seguros de Vida Individuales, analizando el impacto de las cancelaciones en este tipo de seguros; así como las principales estadísticas y causas que determinan dichas anulaciones.

Con los datos recogidos de las estadísticas obtenidas de diversas compañías del sector asegurador[27]; ICEA calcula la Vida Media de la Cartera del ramo de Vida Individual, siendo ésta de 6,2 años en el cierre del ejercicio del 2012[28].

Así mismo, afirma que, a excepción del 2009 (año de crisis financiera), dicha vida media ha sido muy similar dentro de los cinco años anteriores al 2012 *(Tabla 1)*:

| 2012 | 2011 | 2010 | 2009 | 2008 |
|---|---|---|---|---|
| 6,2 años | 6,4 años | 6,5 años | 5,6 años | 6,1 años |

**Tabla 1:** *Histórico de Vida Media de la Cartera de Seguros de Vida Individual*
**Fuente**: *ICEA, 2013*

Con base en la misma información, se puede determinar la evolución de la retención a lo largo de últimos 10 años, observando una considerable disminución del índice de retención, lo cual se traduce en un incremento en las Tasas de Caída en Pólizas año tras año *(Figura 3)*:

[26] Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones, institución dedicada a realizar trabajos de Investigación sobres temas relacionados con la práctica aseguradora, con el objetivo de analizar tendencias y comportamientos de mercado.
[27] Cabe mencionar que la muestra que ICEA ha realizado se toma a partir de entidades aseguradoras que operan dentro de territorio español.
[28] Dentro del Informe n° 1309- Caída en el Ramo de Vida. Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones (ICEA).

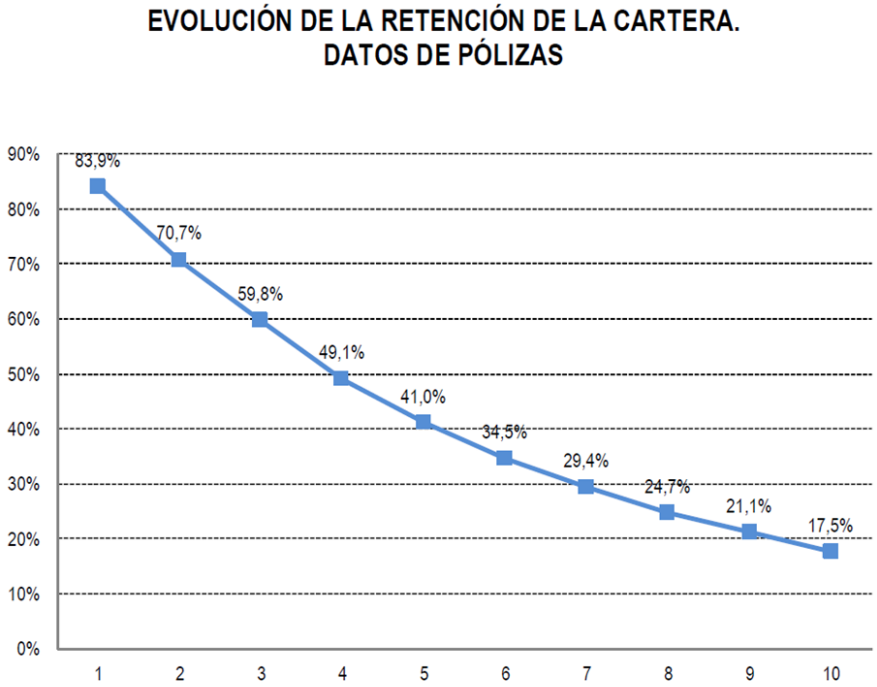


**Figura 3:** *Gráfico de la Evolución de la Retención de Cartera*
**Fuente**: *ICEA, 2013*

Como último dato obtenido a partir de dicha estadística, ICEA muestra las Tasas de Caída de Cartera al cierre del ejercicio 2012, en términos de N° Pólizas y por Importe de Primas, como ya habíamos comentado, con el fin de mostrar una visión económica de la caída de volumen de negocio que representa para las entidades *(Tabla 2)*:

**ÍNDICES DE CAÍDA DE CARTERA EN EL AÑO 2012**

| Productos | Pólizas | Primas Periódicas |
|---|---|---|
| Riesgo | 16,96% | 15,13% |
| Planes Prev. Asegurados | 9,23% | 4,69% |
| Seg. Ahorro/Jubilación | 15,18% | 10,03% |
| **Total Vida Individual** | **16,06%** | **10,52%** |

**Tabla 2:** *Tasas de Caída de Cartera por Tipo de Producto al cierre del 2012*
**Fuente**: *ICEA, 2013*

Ahora bien, como ya también se ha dicho, ha de considerarse la necesidad de estudiar los factores por las que un asegurado decide anular su contrato de seguros. Para ello, cabe mencionar primeramente que la caída de cartera presenta dos tipos de procedencia, a efectos de su análisis y previo a su clasificación (Milán Aguilar y Muñoz Colomina, 2001):

- Voluntaria: Cuando por razones técnicas o características del producto, se llega al vencimiento de la póliza o se decide anular la cartera de un tipo de producto
- Involuntaria: Cuando es el propio cliente el que toma la decisión de abandonar la compañía por razones que considere oportunas

Teniendo presente la procedencia del conjunto de causas registradas por las entidades por las que se producen las anulaciones; resulta necesario revisar y analizar la clasificación de dichas causas. Por lo general, se atribuyen a cuatro tipos de anulaciones de acuerdo a las razones que las motivan (Informe "Caída en el Ramo de Vida. Estadística año 2013" publicado por ICEA); mismas que se han considerado con el fin de ser congruentes con la clasificación realizada por las estadísticas realizadas por ICEA:

- **Vencimiento o Expiración Natural:** Esto sucede cuando la anulación se produce de forma natural por la propia desaparición del riesgo. Por ejemplo, cuando se trata de un seguro temporal y éste ha llegado a su fecha fin de contrato; o bien se ha llegado a la fecha de jubilación dentro de un Seguro de Jubilación o Retiro.

- **Siniestro:** En este caso, es cuando por razones técnicas se ha dado la cancelación. Es decir, cuando ha acaecido el riesgo asegurado (fallecimiento, invalidez) y por lo tanto, se da por finalizado el contrato de seguros.

- **Rescate:** Es similar al anterior, ya que la anulación se produce por razones técnicas; sin embargo, se debe diferenciar ya que en este caso, el evento asegurado no ha sucedido pero el cliente ha decidido retirar, parcial o totalmente, el importe correspondiente a la provisión matemática constituida sobre el riesgo contratado; y por lo tanto, queda la póliza automáticamente dada de baja.

- **Sin efecto o Impago:** Se trata de contratos de seguros que se cancelan por el impago del importe de primas; o bien se consideran “sin efecto” al producirse por el reemplazo, es decir se emite una nueva póliza con alguna modificación realizada.

De esta manera, cabe mencionar que es de vital importancia el correcto registro de dicha clasificación, ya que la depuración y calidad de los datos determinará,

en gran medida, la robustez e implicaciones de los resultados obtenidos ante cualquier análisis realizado a partir de esta información.

Con base en ello y de acuerdo al análisis realizado por ICEA, se puede ver el peso que toma cada una de dichas causas dentro las anulaciones globales que registran las entidades dentro su cartera de clientes mediante el siguiente gráfico *(Figura 4)*:

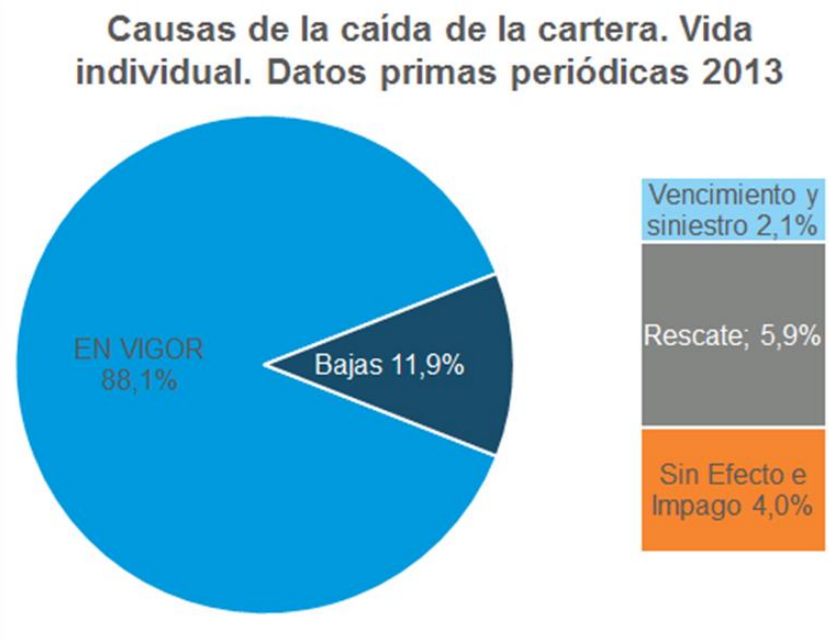


**Figura 4:** *Distribución de Caída de Cartera por Causas*
**Fuente**: *Informe "Caída en el Ramo de Vida. Estadística año 2013" publicado por ICEA*

# CAPITULO 3: TRATAMIENTO DE LA INFORMACIÓN

## 3.1. Introducción

Antes de iniciar con cualquier tipo de aplicación empírica, se debe dedicar un apartado al tratamiento de la información para presentar el conjunto de datos y calidad de la información con la que se cuenta. Saber leer e interpretar los datos que utilizamos es importante para su posterior manejo y correcta interpretación de la información.

Existen muchas formas de presentar y analizar la información; ya sea por medio de tablas de datos o haciendo uso en algunas ocasiones de ciertos gráficos que ayuden a ser mucha más visible su comprensión. Sin embargo, el objetivo principal del tratamiento de las bases de datos radica en la interpretación de los datos que se están manipulando; ya que de aquí dependerá gran parte de las conclusiones que se obtengan a partir de dicha información.

Siendo así, se ha decidido dedicar este capítulo a la generación y análisis de la información con la que se cuenta; mediante la descripción del *contexto, muestra y variables* con las que se desarrollaran las aplicaciones empíricas presentadas en capítulos posteriores. Para ello, se hará una breve reseña del contexto global sobre el que se encuentra el Seguro de Vida dentro del sector asegurador español. Esto seguido del análisis del conjunto de variables consideradas mediante la presentación de cierta estadística descriptiva de los datos. Así mismo, se hará mención a las características generales de la muestra con la que se parte; todo ello con el fin de conocer con mayor detalle el tipo de información con la que se trabajará.

### 3.2. Contexto

El Seguro de Vida ha presentado un progreso paulatino y más pausado que el negocio de los Seguros de No Vida. Aun así, poco a poco se ha ido fortaleciendo y posicionándose como uno de los recursos fundamentales para el desarrollo económico y social. Sin embargo, desde sus inicios, se ha tenido como un mecanismo de referencia debido a la esencia de sus objetivos: seguridad y solvencia. En otras palabras, se podría aludir la importancia de un Seguro de Vida mediante las bondades que ofrece:

- Contribuir a la formación de un hábito de ahorro seguro planificado y continuado en el tiempo que ofrezca solidez a la economía familiar

- Cubrir las consecuencias económicas derivadas del fallecimiento de una persona que pudiese llevar consigo un estado de intranquilidad o desorden social

- Por el contrario, contribuir a la constitución a un ahorro colectivo con el fin cubrir el evento opuesto, la longevidad de la sociedad, de tal forma que se logren cubrir las necesidades económicas que se pueda generar durante una vida tan larga

- Contribuir a la solidez y desarrollo económico a nivel país, mediante el considerable volumen de reservas técnicas derivadas de los seguros de vida que se direccionan en inversiones sustentables y seguras

Bajo este contexto se asienta la relevancia del Seguro de Vida dentro del marco socio-económico de un país; y por tanto, cuestión de interés para efectos del presente estudio.

De la misma forma, en que lo han tenido a bien considerar otros autores (Martínez, 2012), es interesante complementar dicho contexto mediante el posicionamiento que tiene el negocio de Vida en cuanto a su crecimiento a nivel mundial.

Mediante el siguiente gráfico se puede comparar la variación de primas que presentó el negocio de Vida en el 2013 frente al negocio de No Vida *(Tabla 3):*

| | Vida | No vida | Total |
|---|---|---|---|
| Mercados avanzados | 3,8% | 1,8% | 2,9% |
| Mercados emergentes | 6,9% | 8,0% | 7,4% |
| **Mundo** | **4,3%** | **2,9%** | **3,7%** |

**Tabla 3:** *Variación de Primas en el año 2014*
**Fuente**: *Swiss Re Economic Research & Consulting. (SIGMA Nº4 / 2015 )*

Así se puede observar que el ramo de Vida presenta un crecimiento mayor frente al ramo de No Vida, con excepción de los Países Emergentes. Lo cual, hace suponer que, a pesar de la evolución lenta que ha caracterizado al ramo de Vida, empieza a cobrar importancia en la medida en la que la conciencia y cultura del seguro y protección toma fuerza.

Ahora bien, a pesar de este crecimiento generalizado en primas en el 2014 en el seguro de Vida; los mercados avanzados se mantienen en un estancamiento desde la crisis económica del 2008.

Por el contrario, en los mercados emergentes, el crecimiento en primas ha sido más lento después de la crisis *(Figura 5)*:

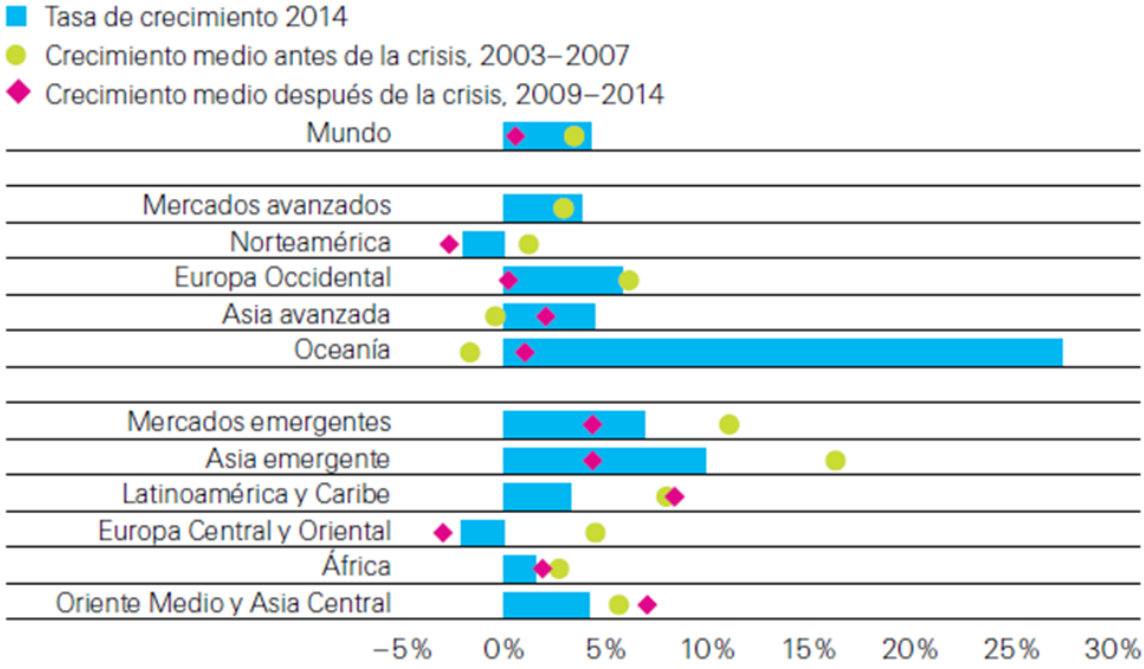


**Figura 5:** *Crecimiento de Primas del Ramo de Vida antes y después de la crisis económica del 2008*
**Fuente**: *Swiss Re Economic Research & Consulting. (SIGMA Nº4 / 2015 )*

Sin embargo, las primas de los mercados avanzados muestran un crecimiento mucho más acelerado que el crecimiento de la economía. Esto es, las primas del Ramo de Vida en los mercados avanzados crecieron un 3,8% en el 2014 *(Tabla 3)*; siendo este crecimiento mejor al crecimiento que presenta el PIB.

Observando un aumento en la penetración del seguro, se tiene que el gasto per cápita en seguro de Vida en los mercados avanzados aumentó en el año 2014; situación que comparte el negocio asegurador del ramo de Vida en España junto con el resto de países de la región *(Figura 6).*

Desde esta perspectiva se observa que, si bien existe un crecimiento generalizado que presenta el grupo de Mercados Desarrollados en el negocio de Vida; dicho crecimiento es lento. De ahí que cobre importancia la realización de estudios como el presente, donde se analiza y ofrecen alternativas de control del riesgo de caída de cartera o retención de clientes para ayudar a acelerar dicho crecimiento.

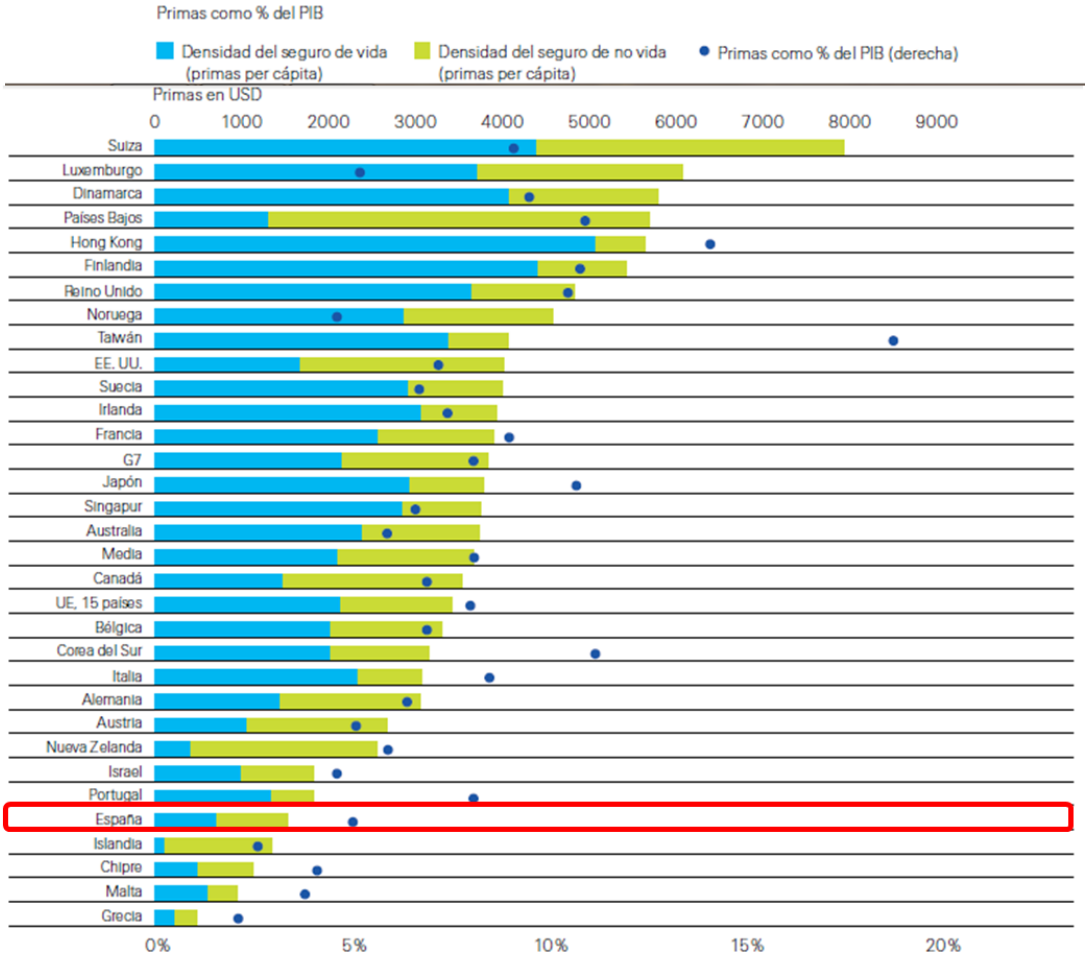


**Figura 6:** *Densidad y penetración del seguro en los mercados avanzados en el 2014*
**Fuente**: *Swiss Re Economic Research & Consulting. (SIGMA Nº4 / 2015 )*

### 3.3. Muestra

Por otro lado, con el fin de ofrecer mayor detalle e información en cuanto a la muestra global que se ha de utilizar en la aplicación empírica; se debe mencionar que la información se ha obtenido íntegramente de una cartera real de pólizas de seguros pertenecientes a una compañía aseguradora que actualmente opera dentro del mercado asegurador español. Sin embargo, de acuerdo con la *Ley de Protección de Datos Personales*[29], se obtuvo dicha información sin tener acceso a ningún tipo de dato personal que hiciese referencia al tomador de las pólizas (como nombre, dirección, teléfono, etc.). Es por ello, que se han tenido limitaciones en el acceso a cierta información, o bien a la completa información de ciertas variables que integran la cartera muestra.

No obstante, con el fin de tener en mente las dimensiones y tipo de cartera a la que se ha tenido acceso para la aplicación de este estudio; se puede concluir esta sección con una breve estadística general de la muestra *(Tabla 4 y Figura 7).*

| RESUMEN DESCRIPTIVO DE CARTERA | |
|---|---|
| Nº TOTAL DE POLIZAS | 19,784 |
| EDAD MEDIA CARTERA | 47 |
| DESVIACION TIPICA EDAD CARTERA | 12.47 |
| MODA EDAD CARTERA | 47 |
| MEDIANA EDAD CARTERA | 47 |
| EDAD MEDIA HOMBRES | 47 |
| EDAD MEDIA MUJERES | 42 |

**Tabla 4:** *Estadística general de la cartera muestra - Edad*
**Fuente**: *Propia de los autores*

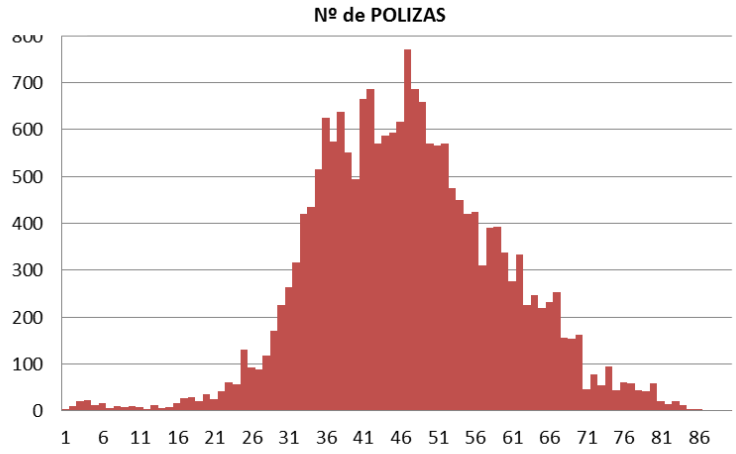


**Figura 7:** *Histograma - EDAD*
**Fuente**: *Propia de los autores*

[29] Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal

Como se puede apreciar, se dispone de una cartera con una Edad Media en torno a los 47 años; la cual coincide con la Moda y Mediana de la muestra. En cuanto a la Edad Media separada por Edad, se observa que para el colectivo de Mujeres, ésta se encuentra por debajo de la media general, sin ser demasiado significativa la diferencia. Cabe mencionar que dicha Edad Media se encuentra en sintonía con el sector asegurador, ya que se considera una edad en la que la toma de decisión o interés asegurable es un tema estable. En otras palabras, una cartera demasiado joven podría dar problemas de variabilidad, o bien una cartera más madura de edad puede verse en complicaciones en cuanto a la rentabilidad que genera.

### 3.4. Variables Utilizadas

La selección de variables en modelos de estimación las tasas de anulación de pólizas dentro de una entidad aseguradora es un tema complicado; ya que pueden existir diversos factores influyentes en la Caída e Cartera. Por un lado, hay factores ligados al propio cliente; como edad, sexo; directamente ligadas a las características de la póliza como son la antigüedad, el tipo de seguro; o bien ligados a al canal de venta o niveles de competencia en que ha sido adquirido el contrato de seguros. Por otro lado, dicha selección se podrá ver limitada por la propia información disponible que se pueda considerar. De aquí la importancia de tener un adecuado control y robustez del registro de las causas de cancelación; así como la calidad de los datos que se tienen dentro de las entidades aseguradoras.

La información obtenida de otros estudios previos sobre el comportamiento de la fidelidad de los asegurados hace referencia a factores sobre la demanda de productos de seguros. Sin embargo, existe poca literatura acerca de la elaboración de escenarios de Caída de Cartera o factores influyentes de la misma. Por mencionar algunos de éstos, se inicia hablando del nivel de ingresos (*Hammond et al. 1967*) como un primer factor influyente en el comportamiento de este riesgo. Posteriormente, no fue hasta los ochenta cuando se empieza a hablar de la retención y fidelización de los clientes mediante estudios del marketing relacional (*Crosby y Stephens, 1987*).

También aparece el tema de la calidad del servicio ofrecido por las entidades y de ahí el nivel de satisfacción de sus clientes (*Wells y Stafford, 1995*). Y más recientemente, se destacan trabajos de estrategias fidelización donde se identifican variables claves como son la antigüedad, la edad y género del asegurado o tipo de cobertura asegurada (*Cooley, 2002*). Todo ello, recogido y considerado en el más reciente estudio del cálculo de escenarios de caída considerando el contagio entre cancelaciones (Ayuso, Guillén y Pérez-Marín; 2011); en el cual, además de tener en cuenta todas estas consideraciones mencionadas, se optó por realizar una segmentación por homogeneidad de productos y antigüedad de la póliza como argumento base para desarrollar su estudio.

Pues bien, como se puede observar, la probabilidad de cancelación de una póliza puede depender o se atribuye a múltiples factores; de ahí que la selección de variables sea un tema complejo. Por otro lado, también se puede ver limitado por la muestra que se tenga disponible; como quizá sea el caso del presente estudio. Sin embargo, servirá como base para la realización de futuros estudios donde se disponga o bien, de mayor volumen de cartera en cuanto a número de pólizas se refiere; o lograr completar el número de variables cualitativas que se tengan registradas en los sistemas informáticos.

Ahora bien, como ya se ha comentado, existen diversas causas asociadas a las razones que motivan la cancelación de la póliza. Sin embargo, podríamos englobarlas en dos tipos que, por la propia naturaleza de su origen, deberían causar distintos impactos. Por un lado, aquellas causas que deberían implicar poco impacto a la entidad, ya que no existe una transferencia del riesgo hacia la competencia. Como es el caso, de las clasificadas por ICEA como *Vencimiento*, *Expiración Natural* o *Siniestro*. Es decir, se da por finalizado de "manera natural" el riesgo cubierto y no debería existir la posibilidad de que el cliente se mueva hacia otra compañía. Por otro lado, se tienen los *Rescates* o *Impago de Primas*, las cuales son las que impactan mayormente a las entidades; ya que en este caso, el cliente es el que ha decidido ya no tener cubierto el riesgo o bien finalizar su relación contractual con la compañía por una mejor oferta del mercado. Es decir, la entidad debe tener en cuenta este evento como posible desviación de sus márgenes futuros. Cabe hacer mención de ello, ya que en el presente análisis no se dispone de la clasificación de la causa de anulación, por lo cual sólo se centrará en diagnosticar el tipo de clientes propenso a la anulación independientemente de la razón que lo conlleva a hacerlo.

Con dicho estado de situación, y con base en la accesibilidad que se ha tenido a los datos utilizados; se ha seleccionado las siguientes variables que son susceptibles a explicar el comportamiento de la tasa de caída que presenta una entidad aseguradora.

A continuación se enlistan y se realiza un breve análisis descriptivo de dichas variables a considerarse en la aplicación empírica que se aborda (*Tabla 5*).

| NOMBRE | DESCRIPCION |
|---|---|
| SEXO | Sexo del asegurado |
| EDAD_ACTUARIAL | Edad del asegurado al cierre del ejercicio |
| ANTIGUEDAD | Años de antigüedad de la póliza desde su Fecha de Emisión hasta la Fecha de Cálculo |
| TIPO_PRODUCTO | Tipo de Producto contratado: Individual o Colectivo; Ahorro o Riesgo |
| TIPO_PRIMA | Tipo de Prima: Unica o Periodica |
| RED | Tipo de Red: Propietaria o No Propietaria |
| FORMA_PAGO | Periodicidad del pago de la prima |
| ANO_EFECTO | Año de Emisión ó Efecto de la póliza |
| EDO_CIVIL | Estado Civil del asegurado, si lo ha comunicado |
| HIJOS | Tiene (ó no) hijos el asegurado, si lo ha comunicado |
| VALOR_CLIENTE | Valor "comercial" definido por la Aseguradora de acuerdo a metodologías internas |
| ICE | Indice de Capacidad Económica |
| NIV_INGRESOS | Nivel de Ingresos |
| NIV_ESTUDIOS | Nivel de Estudios |
| **TIPO_PRESTACION** | Indica si la póliza esta en Vigor o Anulada |

**Tabla 5:** *Variables seleccionadas para la aplicación empírica*
**Fuente**: *Propia de los autores*

- **SEXO**

Comercialmente hablando, se dice que las compañías de seguros suelen cobrar diferentes primas a hombres y a mujeres: “El seguro médico es menos costoso para los hombres jóvenes y de mediana edad”; “Las compañías de seguros de automóviles cobran más por adolescentes varones”; “Los seguros de vida son más caros para los hombres que para las mujeres”..., etc.

Estas afirmaciones suelen ser ciertas debido a las metodologías e hipótesis utilizadas en la tarificación de un Seguro. De hecho, en el ramo de Vida es de vital importancia debido a los factores biométricos que son considerados en el momento de tarificar la prima del seguro. Es por ello que esta variable es muy utilizada y debe ser considerada en la cuantificación de cualquier tipo de riesgo asumido por una entidad aseguradora.

A partir del 21 de diciembre de 2012, las entidades aseguradoras no pueden tener en cuenta el sexo para fijar el precio de sus servicios. Las compañías deben utilizar otras variables para segmentar a sus clientes, como por ejemplo los hábitos y estilos de vida o sus antecedentes, tal como ya se hace en otros países europeos. Está

documentado que dicha Directiva[30] busca aplicar la igualdad de género en el acceso a bienes y servicios ya que prohíbe la discriminación por razón de sexo.

Sin embargo, además de la polémica que ha causado dicha Directiva, se sabe que el hecho de tener en cuenta el sexo como factor de cálculo, no es cuestión de discriminación como tal; sino que existen estadísticas que pueden diferenciar justificadamente la evaluación del riesgo en un género con respecto al otro. No obstante, en términos de siniestralidad, dicho comportamiento también muestra diferencias significativas; lo cual nos da razón suficiente para ser considerada esta variable como posible factor explicativo del comportamiento del asegurado propenso a anular de una póliza.

Por lo que, analizando la distribución por SEXO de la cartera muestra se obtiene la siguiente información (*Tabla 6 y Figura 8*):

| SEXO | Descripción | Nº de POLIZAS | % Peso |
|---|---|---|---|
| H | Femenino (Hembra) | 8,054 | *40.71%* |
| V | Masculino (Varón) | 11,730 | *59.29%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 6:** *Distribución de la muestra por la variable SEXO*
**Fuente**: *Propia de los Autores*

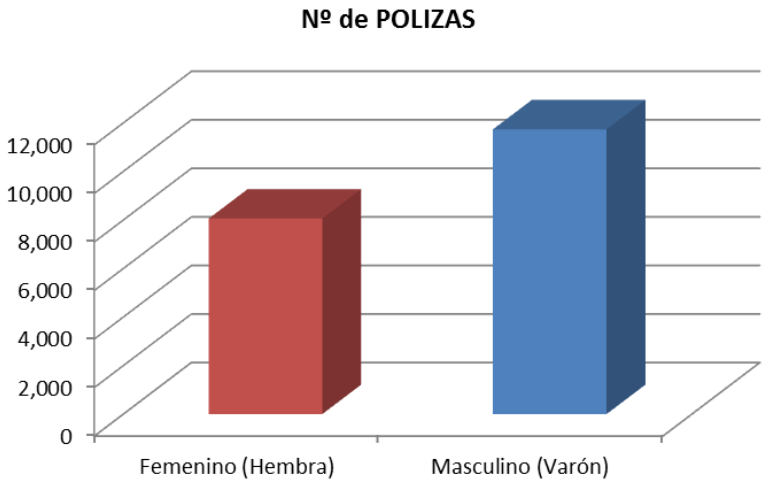


**Figura 8:** *Gráfico de la Distribución por SEXO*
**Fuente**: *Propia de los Autores*

[30] Directiva 2004/113/CE

Como se puede observar, el grupo más representativo en la muestra viene dado por el grupo de los Hombres.


**Figura 9:** *Perfil de Fallecimientos por Sexo*
**Fuente**: *ICEA, Memoria Social del Seguro Español 2012*

Lo cual se trata de una distribución bastante aceptable comparada con el sector dentro de este ramo en términos de siniestralidad. Esto es, en otras palabras, el perfil de los fallecimientos muestra una considerable diferencia entre la tasa de aseguramiento de hombres y mujeres, según información recogida por ICEA (*Figura 9*).

- **EDAD_ACTUARIAL**

Es un concepto esencial que se debe considerar en cualquier estudio de riesgos biométricos; ya que es una variable generalmente utilizada dentro de las técnicas actuariales utilizadas en el ramo de Vida.

En principio, cabe esperar que a mayor edad, exista una mayor tendencia hacia la adquisición de un seguro de Vida; ya que aumenta el interés asegurable y se da mayor prioridad a la tranquilidad y protección que ofrece este tipo de servicio. O bien, si ya se cuenta con una póliza de seguros, también suena lógico pensar que con el aumento de la edad, se supondría una menor intención de anular el contrato dado que es más razonable pensar que es más cercano el momento de utilizar dicho servicio.

Pues bien, esto se puede sustentar tomando las estadísticas obtenidas nuevamente por ICEA, con base en el nivel de esfuerzo que presentan la población española en la adquisición de un seguro (*Figura 10*).

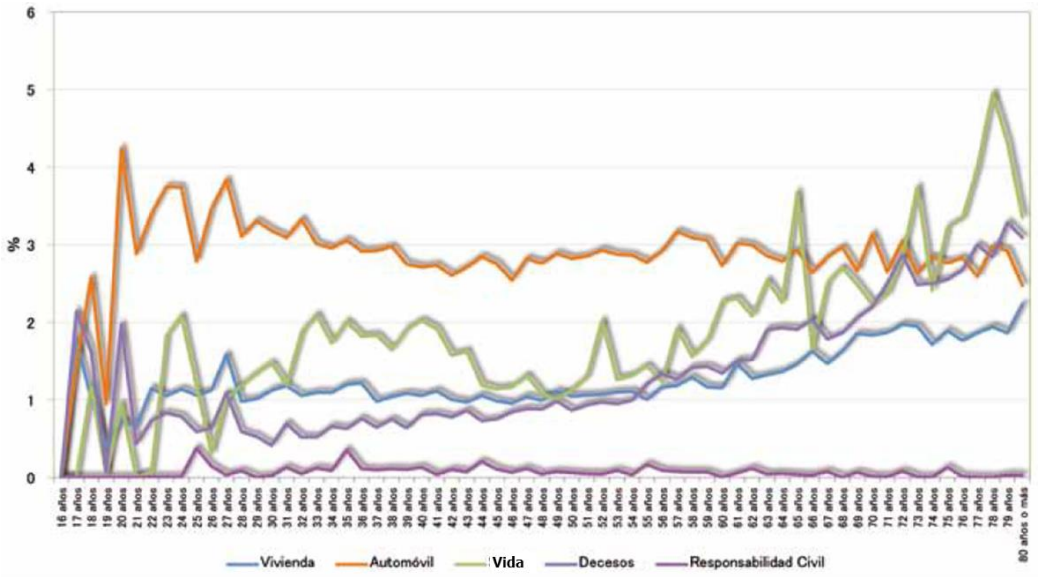


**Figura 10:** *Esfuerzo de los hogares por adquirir seguros, según la edad de su sustentador principal*
**Fuente**: *ICEA, Memoria Social del Seguro Español 2012*

En este gráfico se puede ver la tendencia creciente en el ramo de Vida conforme aumenta la edad. De hecho, dicho interés se atenúa a partir de los 55 años, lo cual parece sugerir que la caída de cartera debe ser menor, o bien estabilizarse, a partir de dicho rango de edad.

Así bien, para efectos de la cartera muestra utilizada para esta aplicación empírica, se tiene la distribución por edades en la siguiente tabla (*Tabla 7 y Figura 11*):

| COD | RANGO DE EDAD | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 1 | [ 0 - 37 ] | 3,865 | *19.54%* |
| 2 | [ 37 - 43 ] | 3,609 | *18.24%* |
| 3 | [ 43 - 49 ] | 3,824 | *19.33%* |
| 4 | [ 49 - 58 ] | 4,444 | *22.46%* |
| 5 | [ 58 - y más ] | 4,042 | *20.43%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

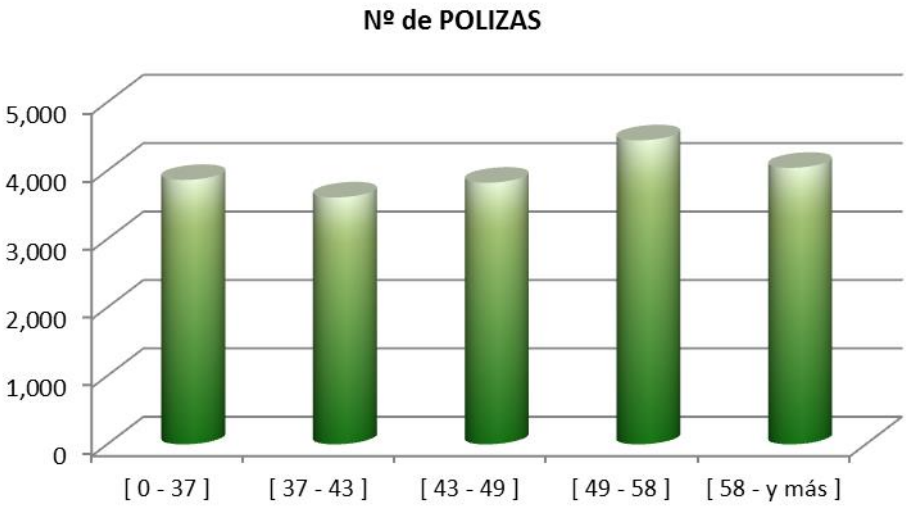


**Figura 11:** *Gráfico de la Distribución por EDAD*
**Fuente**: *Propia de los Autores*

Como se puede observar, el grueso de la muestra se encuentra concentrado en mayores de 40 años. Debido a que la distribución se presenta con base en los códigos asignados a la variable Edad Actuarial (*ver Sección 3.5*), no es sencillo apreciar que la cartera seleccionada muestra la misma tendencia creciente. Sin embargo, excluyendo el rango de [ 0 – 37), ya que es un rango amplio de edad y por ello su peso considerable, la muestra manifiesta la afirmación en cuanto al incremento del interés asegurable conforme avanza la edad del asegurado.

### ❖ ANTIGÜEDAD

Cuando se habla del término antigüedad, es necesario diferenciar entre la antigüedad del cliente, tomando en cuenta todo el conjunto de pólizas que tiene contratada y causando baja cuando todos y cada uno de estos contratos han sido anulados. O bien, la antigüedad propia de la póliza por sí misma. Parece razonable pensar en que cuanto mayor sea el número de pólizas contratadas, mayor fidelidad se tiene por parte del asegurado y por tanto, menor probabilidad de cancelación.

A efectos del presente estudio, se ha considerado la antigüedad de la póliza, ya que no se ha podido tener acceso a alguna variable que agrupe o asocie cada contrato con el cliente. Sin embargo, sí que podremos analizar si el nivel de vinculación que

tiene el asegurado con la compañía influye en la caída de cartera mediante la variable "Valor del Cliente", la cual posteriormente será descrita.

En la siguiente tabla se muestra la distribución de la cartera utilizada para este estudio por años de antigüedad (*Tabla 8 y Figura 12*), donde se puede observar que la mayor parte de la cartera de pólizas utilizadas para este estudio, cuenta con una antigüedad entre 3 y 6 años:

| RANGO DE ANTIGÜEDAD | Nº de POLIZAS | % Peso |
|---|---|---|
| [ 0 - 2 ] | 3,193 | *16.14%* |
| [ 3 - 4 ] | 5,690 | *28.76%* |
| [ 5 - 6 ] | 6,122 | *30.94%* |
| [ 7 - 8 ] | 3,406 | *17.22%* |
| [ 9 - 10 ] | 1,090 | *5.51%* |
| [ 11 - y más ] | 283 | *1.43%* |
| **TOTAL** | **19,784** | ***100.00%*** |

**Tabla 8:** *Distribución de la muestra por la variable ANTIGÜEDAD*
**Fuente**: *Propia de los autores*

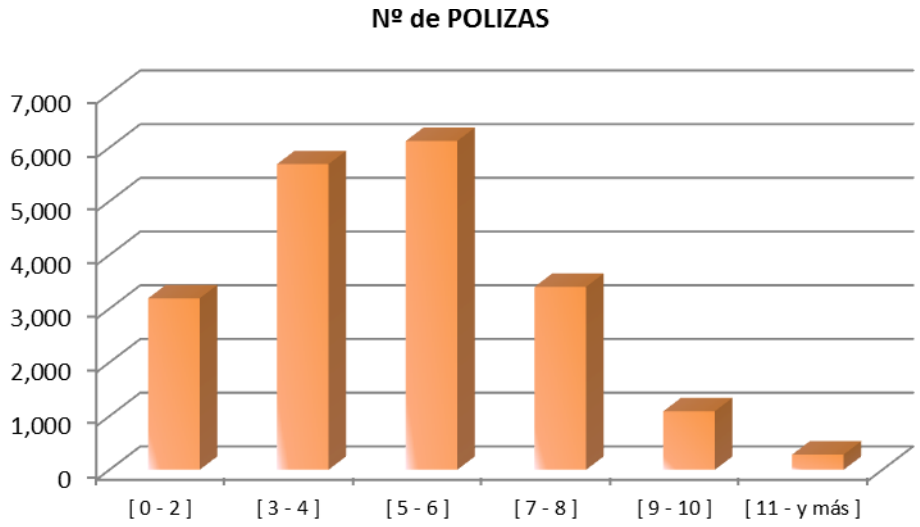


**Figura 12:** *Gráfico de la Distribución por ANTIGÜEDAD*
**Fuente**: *Propia de los Autores*

### ❖ TIPO DE PRODUCTO

La experiencia del sector (de acuerdo a Estadísticas de ICEA) demuestra que existen ramos de seguros en los que las tasas de caída de cartera se comportan de diferente forma. En algunos casos, como es el ramo de Seguros de Automóvil, se ve influenciado por el precio principalmente. En otros, se valora el servicio o bien las

coberturas y cláusulas del contrato por sí mismo. Es por ello, que dentro del propio ramo de Vida, también puede suponer que se pueda dar esta diversidad de acuerdo al tipo de producto contratado.

El Seguro de Vida es una de los tipos del Seguro de Personas en el que el pago por parte de la compañía de seguros de la suma asegurada del contrato, depende del fallecimiento o sobrevivencia del asegurado en un momento determinado. De aquí, una primera clasificación del tipo de póliza de seguro de Vida contratado.

Ahora bien, se conoce que los seguros de Vida pueden clasificarse en función de varias características: Por su duración, pueden ser temporales o vitalicios. Por su tipo de Prima, si la prima es constante entonces son productos a Prima Nivelada o a Prima de Riesgo cuando aumenta en función de la edad del Asegurado. O bien, de acuerdo a la cantidad de asegurados cubiertos por la póliza, pueden ser Individuales o Colectivos.

Pues bien, para efectos del presente estudio, se tiene la siguiente distribución de la muestra por la variable tipo de Producto (*Tabla 9 y Figura 13*):

| TIPO PRODUCTO | Nº de POLIZAS | % Peso |
|---|---|---|
| INDIVIDUAL AHORRO | 12,808 | *64.74%* |
| INDIVIDUAL RIESGO | 6,976 | *35.26%* |
| **TOTAL** | **19,784** | ***100.00%*** |

**Tabla 9:** *Distribución de la muestra por la variable TIPO DE PRODUCTO*
**Fuente**: *Propia de los autores*

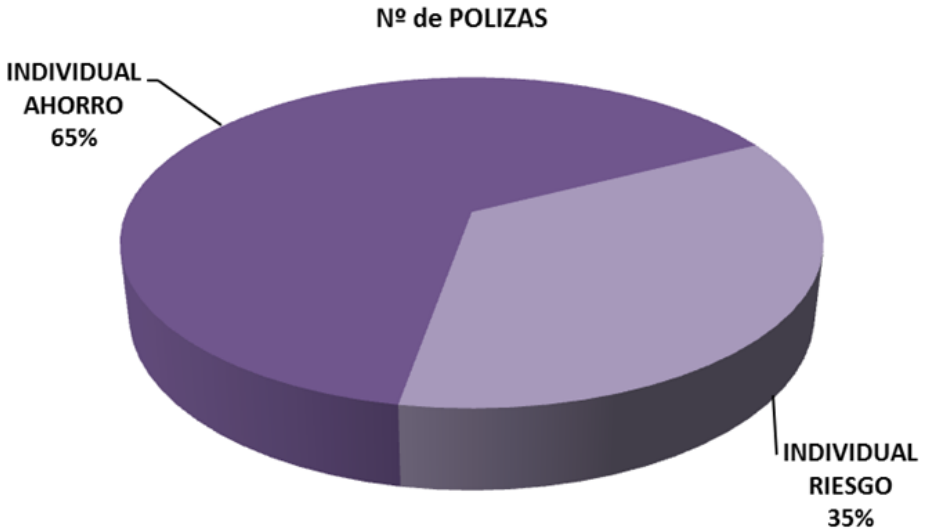


**Figura 13:** *Gráfico de la Distribución por TIPO DE PRODUCTO*
**Fuente**: *Propia de los Autores*

De acuerdo con la información disponible para la aplicación empírica, se ha considerado la clasificación de Tipo de Producto con base en si el pago de la indemnización depende del fallecimiento o sobrevivencia del asegurado, siendo así:

- Individual Ahorro: Comúnmente conocido como Seguro de Sobrevivencia donde el beneficiario (quien generalmente se trata del propio asegurado) tiene garantizado el pago de la indemnización, siempre y cuando no haya fallecido una vez terminado el periodo de vigencia del contrato

- Individual Riesgo: La aseguradora se compromete a realizar el pago acordado a los beneficiarios del seguro tras el fallecimiento del asegurado, ya sea por causa natural o por accidental; pudiéndose producir dicho fallecimiento en cualquier momento desde el inicio de la contratación del seguro

En ambos casos, se trata de Seguros de Vida Individuales, donde sólo existe un asegurado cubierto quien generalmente es el contratante de la póliza.

Ahora bien, a manera de completar el contexto de esta variable, se puede tener en consideración el volumen de asegurados y reservas técnicas que representa el ramo de Vida en sus distintas modalidades *(Tabla 10).*

| MODALIDADES | NÚMERO DE ASEGURADOS | | PROVISIONES TÉCNICAS (€) | |
|---|---|---|---|---|
| | 2.007 | % Δ | 2.007 | % Δ |
| Planes de Previsión Asegurados | 164.450 | *12,4* | 861.909.160 | *23,0* |
| Capital Diferido | 4.951.493 | *-1,3* | 48.356.154.595 | *-0,2* |
| Rentas | 2.847.417 | *0,2* | 65.803.729.343 | *1,6* |
| P.I.A.S. | 177.403 | | 546.530.023 | |
| Vinculados a Activos | 807.934 | *8,1* | 13.770.537.793 | *7,7* |
| **Total Seguros de Ahorro** | **8.948.697** | *2,2* | **129.338.860.913** | *2,1* |
| **Total Seguros de Riesgo** | **23.079.731** | *3,6* | **4.375.456.918** | *9,5* |
| | | | | |
| **TOTAL VIDA** | **32.028.428** | *3,2* | **133.714.317.832** | *2,3* |

**Tabla 10:** *Modalidades del Seguro de Vida en España en el 2007*
**Fuente**: *ICEA. (Jurado Gil, José )*

De manera congruente con la distribución de la cartera muestra que se utiliza para esta aplicación empírica, donde la distribución por la variable TIPO DE PRODUCTO indica que la modalidad de INDIVIDUAL AHORRO es el conjunto más representativo; se observa que en términos de Provisiones Técnicas, se mantiene dicha distribución

dentro del negocio asegurador español. Así mismo, en términos de Asegurados, la modalidad de Seguros de Riesgo muestra un peso relevante dentro del negocio de Vida.

### ❖ TIPO DE PRIMA

Esta variable podría venir explicada con base en el Tipo de Producto, ya que esta puede considerarse como otra forma de clasificar a los contratos de Seguros. En este caso, se trata de catalogar a las pólizas de acuerdo a la forma en que se realiza el pago de su prima.

Así se tiene la distribución de la cartera con base en esta variable de la siguiente forma (*Tabla 11 y Figura 14*):

| TIPO PRIMA | Nº de POLIZAS | % Peso |
|---|---|---|
| PRIMA UNICA | 5,337 | *26.98%* |
| PRIMA PERIODICA | 14,447 | *73.02%* |
| **TOTAL** | **19,784** | ***100.00%*** |

**Tabla 11:** *Distribución de la muestra por la variable TIPO DE PRIMA*
**Fuente**: *Propia de los autores*

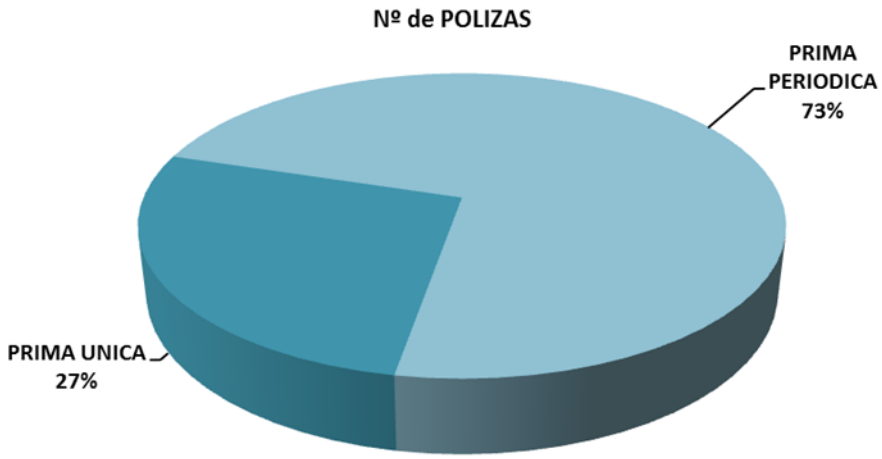


**Figura 14:** *Gráfico de la Distribución por TIPO DE PRIMA*
**Fuente**: *Propia de los Autores*

De esta forma, se ha considerado que la prima puede ser pagada de forma única o periódica. La primera es aquella que se paga de una sola vez, generalmente

asociada a seguros de corta duración. En cambio, la prima periódica es la que se paga, como bien dice su nombre, periódicamente durante la vigencia del seguro.

- **RED**

Hoy en día, existen varios medios por los que se puede adquirir una póliza de seguros; o bien por canales tradicionales como son los agentes y mediadores de seguros; o por los más recientemente creados como lo es el canal de banca-seguros. En ellos recae la labor comercial asesorando y resolviendo cualquier cuestión planteada por el cliente. De aquí la importancia y posible influencia en cualquier movimiento registrado en las pólizas de seguros, como puede ser la anulación de la misma; ya que son el contacto directo con el cliente y que gracias a su experiencia y labor, son una pieza fundamental en temas de captación, fidelización y fuga de la cartera de clientes.

Así se tiene que esta variable hace referencia al canal o red de distribución por el cual se ha realizado la colocación del contrato de seguros. De esta forma, en la base de datos disponible para este estudio, se tiene registrado, si la póliza se ha contratado a través de una Red Propietaria o No Propietaria. Todo ello, englobando y haciendo referencia a las dos figuras de mediadores de seguros que actúan como intermediarios entre las entidades aseguradoras y sus clientes: Agentes de Seguros y Corredores de Seguros.

En términos generales, se puede describir a la figura del Agente de Seguros como aquella persona (física o jurídica) que realiza labores de mediación, promoción, asesoramiento y asistencia con los asegurados o clientes. Lo más destacado es que su vinculación es exclusiva con una entidad aseguradora, a menos que ésta lo autorice expresamente dentro del "contrato de agencia" que pactan. Es así como a partir de un registro de sus agentes, las compañías aseguradoras forman su *Red Propietaria* como un primer canal de distribución de sus productos y servicios.

Por otro lado, se tiene la *Red No Propietaria* que reúne a las figuras de corredores o brokers de seguros que de igual forma son aquellas personas (físicas o

jurídica) que realizan labores similares a las de un Agente pero de forma independiente respecto a cualquier entidad aseguradora y por lo tanto, puede asesorar a los clientes con diferentes ofertas de distintas aseguradoras ya que no tiene ningún vínculo único con alguna entidad específica.

Es así, como se tiene que la distribución de la muestra teniendo en cuenta esta variables es de la siguiente forma (*Tabla 12 y Figura 15*):

| RED | Nº de POLIZAS | % Peso |
|---|---|---|
| PROPIETARIAS (Agentes, Empleados) | 19,015 | *96.11%* |
| NO PROPIETARIAS (Brokers, Corredores) | 769 | *3.89%* |
| **TOTAL** | **19,784** | ***100.00%*** |

**Tabla 12:** *Distribución de la muestra por la variable RED*
**Fuente**: *Propia de los autores*

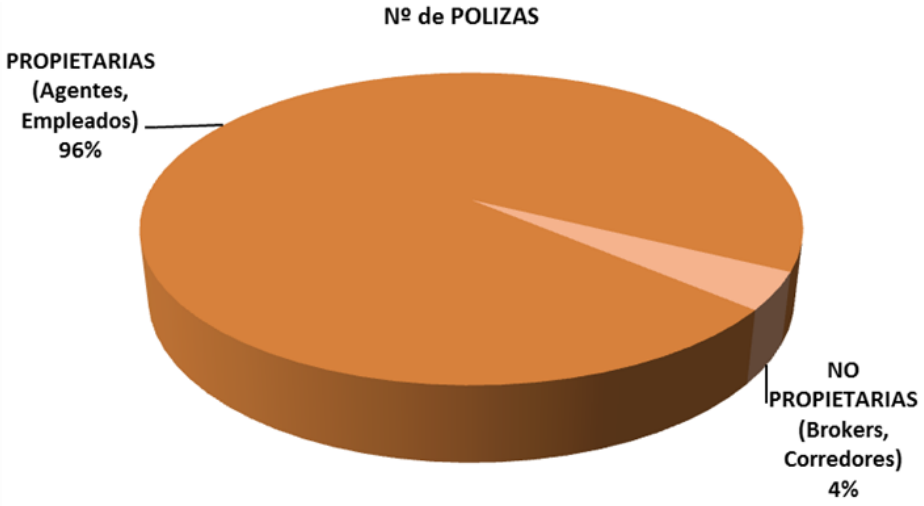


**Figura 15:** *Gráfico de la Distribución por RED*
**Fuente**: *Propia de los Autores*

En el siguiente cuadro, podemos observar un comportamiento similar dentro del sector asegurador español (*Tabla 13*):

**DISTRIBUCIÓN PORCENTUAL DEL SEGUROS DE VIDA POR CANALES***
**(En %)**

| | Primas | | Número de pólizas | |
|---|---|---|---|---|
| | Cartera | Nueva producción | Cartera | Nueva producción |
| Agentes | 14,6 | 15,0 | 23,6 | 19,8 |
| Corredores | 7,7 | 4,3 | 5,0 | 4,0 |
| Bancos y Cajas | 60,0 | 71,1 | 61,3 | 71,0 |
| Oficinas directas | 15,0 | 8,3 | 6,4 | 2,9 |
| Otros | 2,7 | 1,3 | 3,7 | 2,3 |
| **Total** | **100** | **100** | **100** | **100** |

**Tabla 13:** *Distribución del Seguro de Vida por Canales*
**Fuente**: *Dirección General de Seguros, Memoria 2004*

Se puede observar que el mayor peso recae en la Red Propietaria. Esto puede suponerse a que en los procesos de búsqueda y contratación de un seguro, los consumidores prefieren ser asesorados personalmente más que utilizar canales a distancia.

Sin embargo, debido a la característica de "no exclusividad" con la que cuentan los corredores de seguros, puede también pensarse que tiene mayor influencia en el mantenimiento de la póliza en una compañía o su anulación y nueva contratación en otra entidad. De aquí la razón por la cual se ha considerado interesante incorporar esta variable en el estudio.

### ❖ FORMA DE PAGO

Adicional a la variable Tipo de Prima, se cuenta con el registro de esta variable Forma de Pago. A diferencia de la variable Tipo de Prima que está relacionada con el tipo de producto contratado, ésta variable está más vinculada a la prima fraccionada.

Este es un sistema ofrecido por las compañías aseguradoras en determinados ramos de seguros, en donde el asegurado puede abonar la prima de una anualidad completa de forma anticipada, en una sola exhibición o bien ser liquidada en varios pagos periódicos. Sin embargo, esto no significa que el asegurado pueda rescindir del contrato en dichos periodos, sino podrá hacerlo al vencimiento de la anualidad y pagar las primas pendientes hasta dicho vencimiento. Es decir, es sólo una facilidad de pago creada por las entidades aseguradoras para sus clientes.

Bajo esta descripción, la cartera muestra para este estudio presenta la siguiente distribución de acuerdo a esta variable (*Tabla 14 y Figura 16*):

| COD | FORMA PAGO | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 1 | ANUAL | 5,558 | *28.09%* |
| 2 | SEMESTRAL | 1,461 | *7.38%* |
| 3 | TRIMESTRAL | 1,271 | *6.42%* |
| 4 | BIMESTRAL | 0 | *0.00%* |
| 5 | MENSUAL | 6,157 | *31.12%* |
| 6 | UNICA | 5,337 | *26.98%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 14:** *Distribución de la muestra por la variable FORMA DE PAGO*
**Fuente**: *Propia de los autores*

De esta forma, se puede observar que las Formas de Pago más representativas a considerar dentro del estudio, serán las pólizas Anuales, Mensuales y de pago Único.

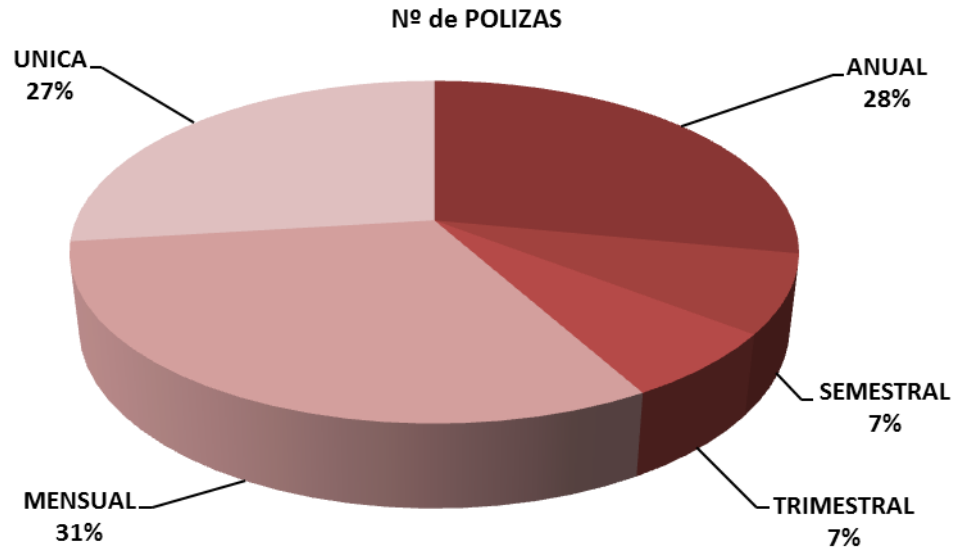


**Figura 16:** *Gráfico de la Distribución por FORMA DE PAGO*
**Fuente**: *Propia de los Autores*

- **AÑO EFECTO**

Como su nombre lo dice se trata de una variable que hace referencia al año en el cual se emitió una póliza de seguros. Desde el punto de vista del volumen de negocio de una compañía de seguros, esta variable toma cierta importancia, al ser una magnitud para medir la evolución y crecimiento de sus operaciones o resultados, que generalmente coincide con el ejercicio contable.

Para los efectos del presente estudio, se ha considerado importante tomar esta variable debido a su posible vinculación con años de crisis o momentos de recesión económica que afecten la producción o motiven la caída de cartera de una entidad aseguradora. Es decir, se podría suponer que la crisis económica del año 2008, puede verse reflejada en los niveles de anulación que presenta el sector asegurador.

Así, se tiene la siguiente distribución de la muestra para este estudio bajo dicha variable (*Tabla 15 y Figura 17*):

| AÑO EFECTO | Nº de POLIZAS | % Peso |
|---|---|---|
| 2002 | 101 | *0.51%* |
| 2003 | 394 | *1.99%* |
| 2004 | 732 | *3.70%* |
| 2005 | 850 | *4.30%* |
| 2006 | 2,263 | *11.44%* |
| 2007 | 2,436 | *12.31%* |
| 2008 | 4,292 | *21.69%* |
| 2009 | 4,761 | *24.06%* |
| 2010 | 3,955 | *19.99%* |
| **TOTAL** | **19,784** | ***100.00%*** |

**Tabla 15:** *Distribución de la muestra por la variable AÑO EFECTO*
**Fuente**: *Propia de los autores*

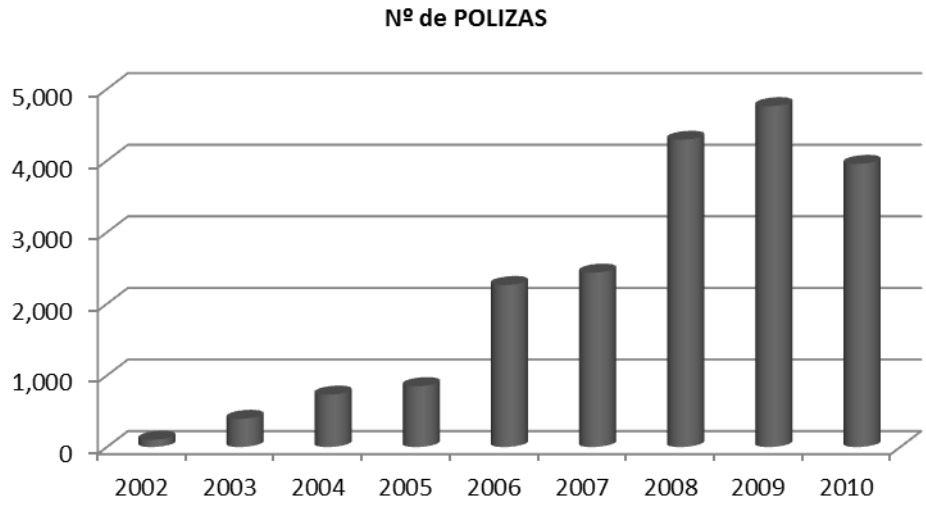


**Figura 17:** *Gráfico de la Distribución por AÑO EFECTO*
**Fuente**: *Propia de los Autores*

Se puede observar que la cartera muestra se encuentra cargada hacia años más recientes. Sin embargo, también se podría intuir que parte de los efectos de la crisis del 2008, se empiezan a reflejar en la bajada de la producción del año 2010. Es así como esta variable podría jugar un papel importante en el patrón de comportamiento de anulación de los contratos de seguros dentro del sector, debido a su posible vinculación con años de recesión económica.

❖ **ESTADO CIVIL**

La composición familiar suele determina diferencias en los índices de consumo, y el sector asegurador no puede ser la excepción. De aquí la razón de considerar esta variable dentro del estudio como posible factor de propensión a la cancelación de una póliza de seguros, ya que cabría suponer que el interés asegurable de una persona puede verse influenciado por el Estado Civil en el que se encuentra o bien variar dependiendo del cambio de dicha condición.

Siendo así, se tiene la siguiente distribución de la muestra según esta variable (*Tabla 16 y Figura 18*):

| COD | ESTADO CIVIL | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 1 | Soltero | 2,422 | *12.24%* |
| 2 | Casado | 15,785 | *79.79%* |
| 3 | Divorciado | 672 | *3.40%* |
| 4 | Viudo | 576 | *2.91%* |
| 5 | Separado | 182 | *0.92%* |
| 6 | Pareja de Hecho | 116 | *0.59%* |
| 9 | Sin Informar | 31 | *0.16%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 16:** *Distribución de la muestra por la variable ESTADO CIVIL*
**Fuente**: *Propia de los autores*

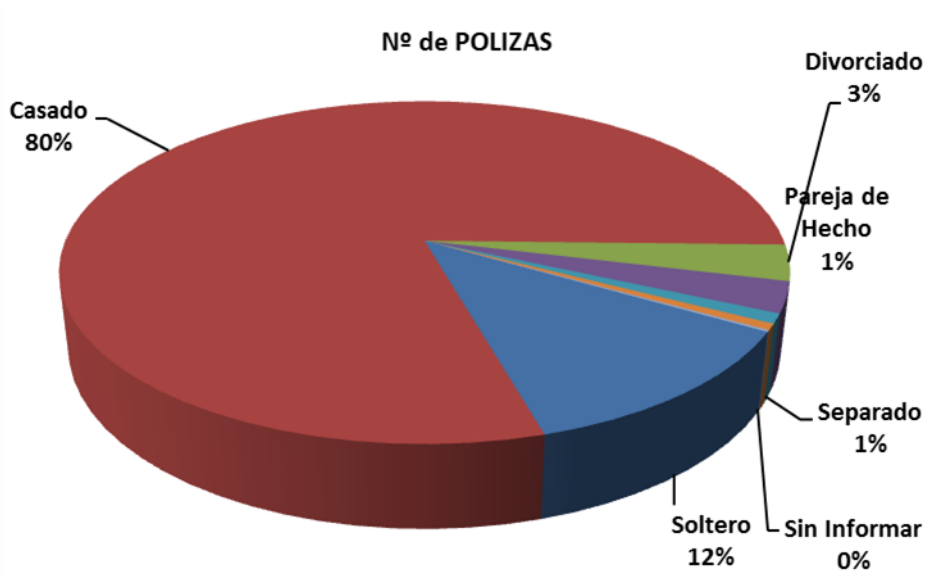


**Figura 18:** *Gráfico de la Distribución por ESTADO CIVIL*
**Fuente**: *Propia de los Autores*

En el siguiente cuadro, se puede observar que esta composición de la cartera coincide con los niveles de penetración de los diferentes tipos de seguros, según el estado civil del sustentador principal (*Tabla 17*), dentro del sector asegurador. En este caso, se debe considerar el ramo de Salud, como el tipo de seguro referencia para el comparativo de la muestra utilizada. Siendo así, se distingue el grupo de "Casado" como el tipo de asegurado con mayor peso, similar a la cartera muestra en estudio.

Tasas de penetración de los diferentes seguros, según el estado civil del sustentador principal.

| | AUTOS | SALUD | VIVIENDA | DECESOS | RC |
|---|---|---|---|---|---|
| Soltero | 72,4% | 23,3% | 64,5% | 33,7% | 9,5% |
| Casado | 90,2% | 32,9% | 81,3% | 55,4% | 14,2% |
| Viudo | 46,4% | 20,3% | 70,0% | 63,8% | 8,2% |
| Separado | 67,1% | 22,5% | 62,7% | 48,7% | 9,4% |
| Divorciado | 73,6% | 25,8% | 68,2% | 43,7% | 9,1% |

**Tabla 17:** *Tasas de Penetración según el estado civil*
**Fuente**: *ICEA, Memoria Social del Seguro Español 2013*

Esto puede deberse al hecho de que la situación de la unidad familiar es importante a la hora de definir la mayor frecuencia del gasto destinado al tema de seguros. Sin embargo, ya existen diferencias significativas entre la situación de casado y la de separado o divorciado; lo cual puede ser importante ser considerado en la influencia que pueda tener en la cancelación o conservación del contrato de seguros. De aquí el por qué se ha tenido a bien considerar esta variable en el estudio.

- **HIJOS**

Esta variable hace referencia a si el asegurado o tomador de la póliza de seguros, tiene o no tiene hijos. Los motivos que sugieren tener en cuenta esta variable son muy similares a la anterior. Cabría suponer que el nivel de aseguramiento de los hogares con hijos puede ser diferente a los hogares donde no hay hijos. Esto podría sugerir que sucede lo mismo con los niveles de anulación de pólizas de seguros que se tengan contratados. Todo ello ligado nuevamente al interés asegurable, ya que este

puede verse influido por la propia economía e intereses familiares que puedan derivarse de la estructura doméstica que tenga los clientes.

Así se tiene la siguiente distribución de la cartera muestra con base en esta variable (*Tabla 18 y Figura 19*):

| COD | HIJOS | Nº de POLIZAS | % Peso |
|---|---|---|---|
| N | Sin hijos | 11,755 | *59.42%* |
| S | Con hijos | 8,029 | *40.58%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 18:** *Distribución de la muestra por la variable HIJOS*
**Fuente**: *Propia de los autores*

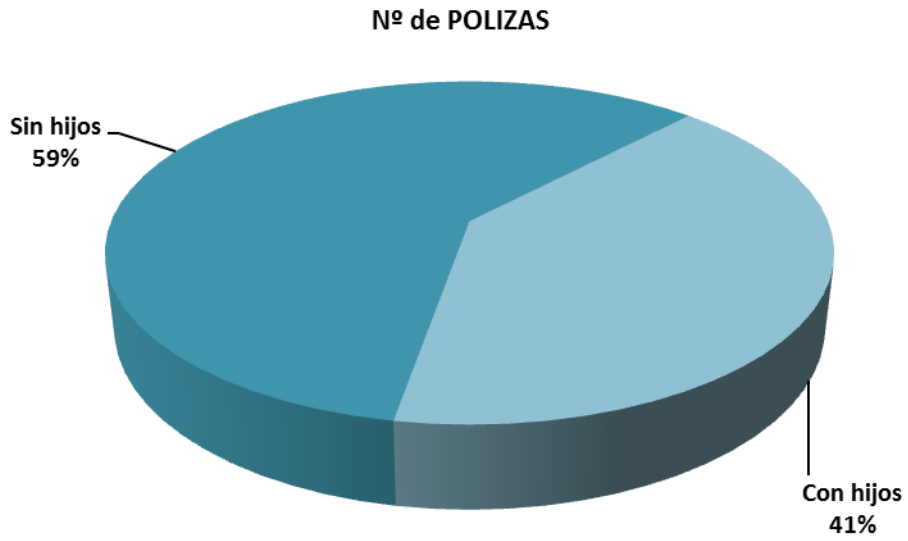


**Figura 19:** *Gráfico de la Distribución por HIJOS*
**Fuente**: *Propia de los Autores*

Se puede observar que la muestra se encuentra bastante equilibrada con respecto a esta variable, predominando el conjunto de asegurados que no tienen hijos por un porcentaje mínimo.

### ❖ VALOR DEL CLIENTE

Esta variable es un concepto creado en su totalidad por la compañía aseguradora proveedora de la información de la muestra; es decir, es una asignación propia de la entidad para clasificar sus clientes. Su principal finalidad es la de lograr

dirigir estrategias de retención sobre aquellos clientes que registren mayor valor (beneficio) para la compañía.

Sin embargo, como su nombre lo dice, es el valor que la compañía le asigna a cada cliente con base en ciertos ratios de siniestralidad, indicadores de rentabilidad y niveles de vinculación que presenta el cliente. Sin ánimo de ser demasiado exhaustivo en la explicación de las metodologías utilizadas para dicha asignación, se puede resumir en el valor que aporta cada cliente considerando tres conceptos: 1) su alta / baja índices de siniestralidad, 2) cómo de rentable son sus pólizas para la entidad y 3) midiendo el nivel de vinculación mediante el número de pólizas contratadas con la compañía. Este tercer punto es el que hace suponer que puede ser un factor de cancelación, ya que cabría pensar que cuanto mayor sea el número de pólizas contratadas por un cliente mayor es su grado de fidelización y por lo tanto, desciende la probabilidad de anulación.

Es así, como se obtiene la siguiente clasificación y distribución de la cartera con base en esta variable (*Tabla 19 y Figura 20*):

| COD | VALOR DE CLIENTE | Nº de POLIZAS | % Peso |
|---|---|---|---|
| A | Muy Rentables y Muy Vinculados | 6,202 | *31.35%* |
| B | Muy Rentables y Medianamente Vinculados | 977 | *4.94%* |
| C | Muy Rentables y No Vinculados | 1,630 | *8.24%* |
| D | Medianamente Rentables y Vinculados | 5,087 | *25.71%* |
| E | Medianamente Rentables y No Vinculados | 1,150 | *5.81%* |
| F | No Rentables y Vinculados | 3,013 | *15.23%* |
| G | No Rentables y No Vinculados | 1,725 | *8.72%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 19:** *Distribución de la muestra por la variable VALOR DEL CLIENTE*
**Fuente**: *Propia de los autores*

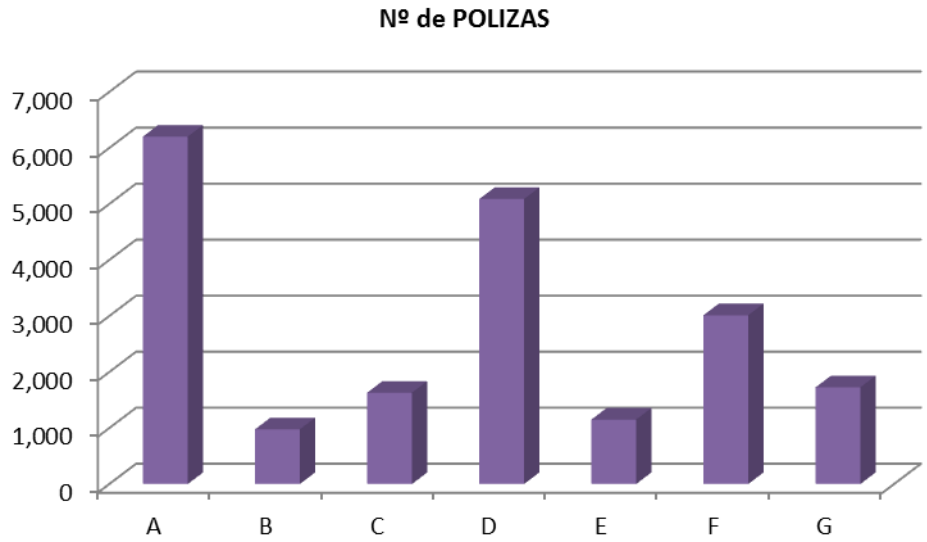


**Figura 20:** *Gráfico de la Distribución por VALOR DEL CLIENTE*
**Fuente**: *Propia de los Autores*

❖ **INDICE DE CAPACIDAD ECONOMICA**

Esta variable, como su nombre lo indica, hace referencia a los niveles de capacidad financiera con la que cuenta el asegurado. Cabe mencionar que, con información de este tipo, se intenta tomar características muy particulares del tipo de cliente propenso a la anulación de la póliza. Sin embargo, gran parte dela problemática de este tipo de variables es que, en la mayoría de las ocasiones, no se logra registrar toda esta información ya que no es proporcionada por todos los clientes.

Sin embargo, aún con la limitación que se puede encontrar en la extracción de este tipo de información, se ha tenido en cuenta esta variable ya que cabría suponer que cuanto mayor sea la capacidad económica del cliente, mayor número de pólizas podría tener contratadas y por tanto, menor probabilidad de fuga podría registrar la compañía ante este tipo de clientes.

Así se tiene que la cartera muestra utilizada para este estudio, se distribuye con base en esta variable de la siguiente forma (*Tabla 20 y Figura 21*):

| COD | ICE | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 1 | Bajo | 1,865 | *9.43%* |
| 2 | Medio-Bajo | 2,688 | *13.59%* |
| 3 | Medio | 6,044 | *30.55%* |
| 4 | Medio-Alto | 6,774 | *34.24%* |
| 5 | Alto | 2,413 | *12.20%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 20:** *Distribución de la muestra por la variable ICE*
**Fuente**: *Propia de los autores*

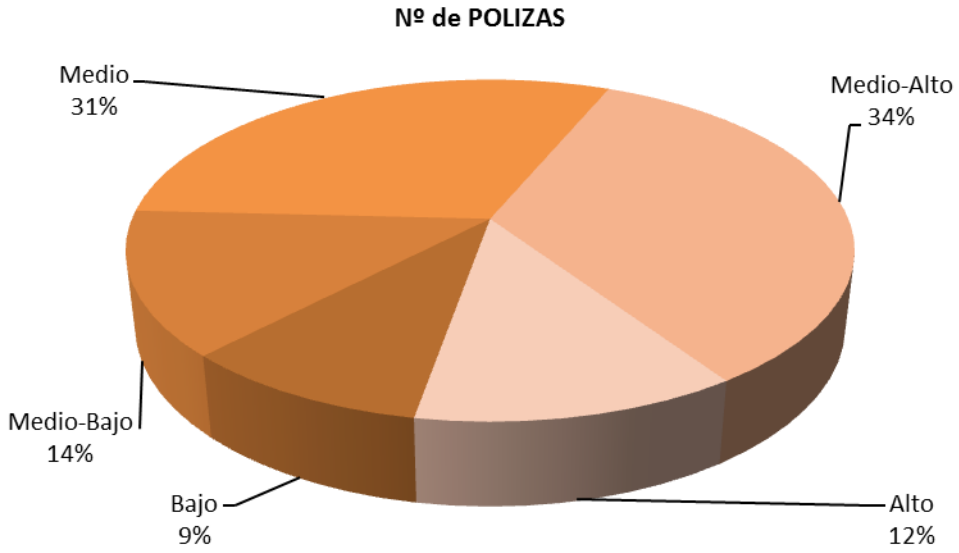


**Figura 21:** *Gráfico de la Distribución por ICE*
**Fuente**: *Propia de los Autores*

❖ **NIVEL DE INGRESOS**

Similar a la variable anterior, se trata de información personal del asegurado, que en la mayoría de las veces, no es posible obtenerla. Este es el caso, de esta variable donde se recoge, como su nombre lo dice, el nivel de ingresos con el que cuenta el cliente que contrata la póliza de seguros.

La distribución de la cartera muestra con base en esta variable es la siguiente (*Tabla 21 y Figura 22*):

| COD | NIV_INGRESOS | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 0 | NO INFORMADO | 1,053 | *5.32%* |
| 1 | < 6.000 | 160 | *0.81%* |
| 2 | 6.000 A 18.000 | 846 | *4.28%* |
| 3 | 18.001 A 36.000 | 915 | *4.62%* |
| 4 | 36.001 A 60.000 | 212 | *1.07%* |
| 5 | 60.001 A 100.000 | 68 | *0.34%* |
| 9 | NS / NC | 16,530 | *83.55%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 21:** *Distribución de la muestra por la variable NIVEL DE INGRESOS*
**Fuente**: *Propia de los autores*

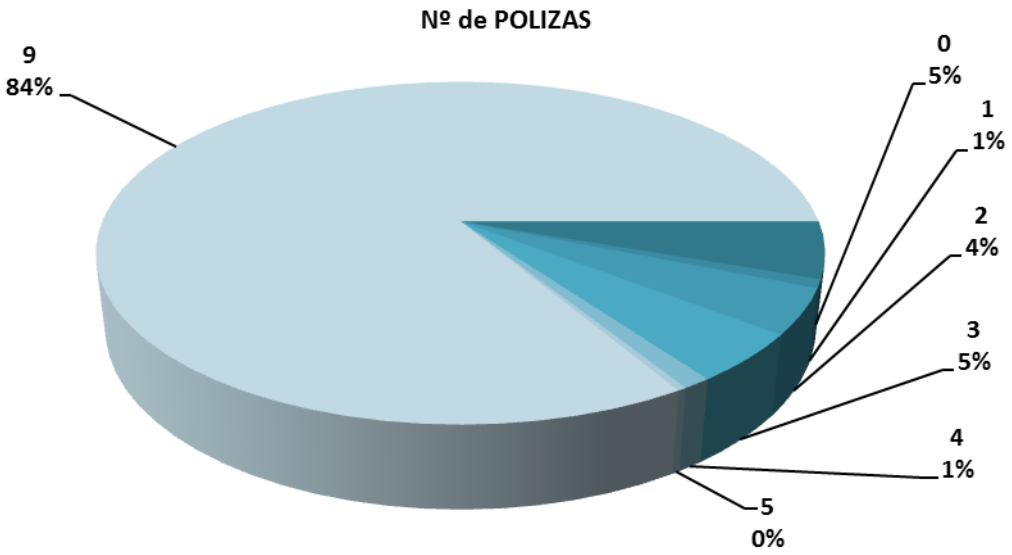


**Figura 22:** *Gráfico de la Distribución por NIVEL DE INGRESOS*
**Fuente**: *Propia de los Autores*

Como ya se ha comentado, es difícil el registro de este tipo de datos; tal es el caso de la cartera utilizada para este estudio. Es por ello, que la mayor parte de la cartera se registra bajo el COD=9 (No se sabe / No se conoce); lo cual deberá tenerse en cuenta en lo momento de analizar los resultados y conclusiones obtenidas.

Ahora bien, obteniendo información del sector, se tiene proporciones similares; donde existe mayor frecuencia en el consumo o gasto en seguros, en tramos de ingresos superiores; es decir en niveles de ingresos medio, medio-alto y alto (*Tabla 22*):

Frecuencia de hogares que gastan en diferentes seguros, según su tramo de ingresos.

| | AUTOS | SALUD | VIVIENDA | DECESOS | RC |
|---|---|---|---|---|---|
| Menos de 500 € | 63,8% | 20,1% | 49,1% | 52,0% | 10,3% |
| De 500 a 1.000 € | 67,3% | 21,1% | 63,8% | 58,8% | 11,4% |
| De 1.000 a 1.500 € | 85,3% | 26,5% | 79,9% | 53,3% | 11,4% |
| De 1.500 a 2.000 € | 91,8% | 34,7% | 87,3% | 47,0% | 12,1% |
| De 2.000 a 2.500 € | 93,9% | 41,1% | 90,6% | 43,9% | 12,8% |
| De 2.500 a 3.000 € | 93,9% | 46,0% | 91,4% | 40,1% | 15,8% |
| 3.000 € o más | 93,8% | 56,0% | 92,2% | 38,0% | 16,9% |
| Total | 81% | 29% | 76% | 52% | 12% |

**Tabla 22:** *Frecuencias de hogares que gastan en seguros según nivel de ingresos*
**Fuente**: *ICEA, Memoria Social del Seguro Español 2013*

### ❖ NIVEL DE ESTUDIOS

Una vez más, se trata de información personal del asegurado, siendo esta variable la que indica el nivel de estudios que ha sido conseguido por el asegurado que adquiere la póliza de seguros. Sin embargo, nuevamente nos enfrentamos a la dificultad del registro de este tipo de información ya que no todos los clientes declaran dicho nivel en el momento de la contratación del seguro.

Así se tiene que la cartera muestra para estudio se distribuye de la siguiente forma con base en esta variable (*Tabla 23 y Figura 23*):

| COD | NIV_ESTUDIOS | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 01 | ELEMENTALES | 626 | *3.16%* |
| 02 | BUP/PP/ESO | 1,497 | *7.57%* |
| 03 | UNIVERSITARIOS | 1,754 | *8.87%* |
| 99 | SIN INFORMAR | 15,907 | *80.40%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 23:** *Distribución de la muestra por la variable NIVEL DE ESTUDIOS*
**Fuente**: *Propia de los autores*

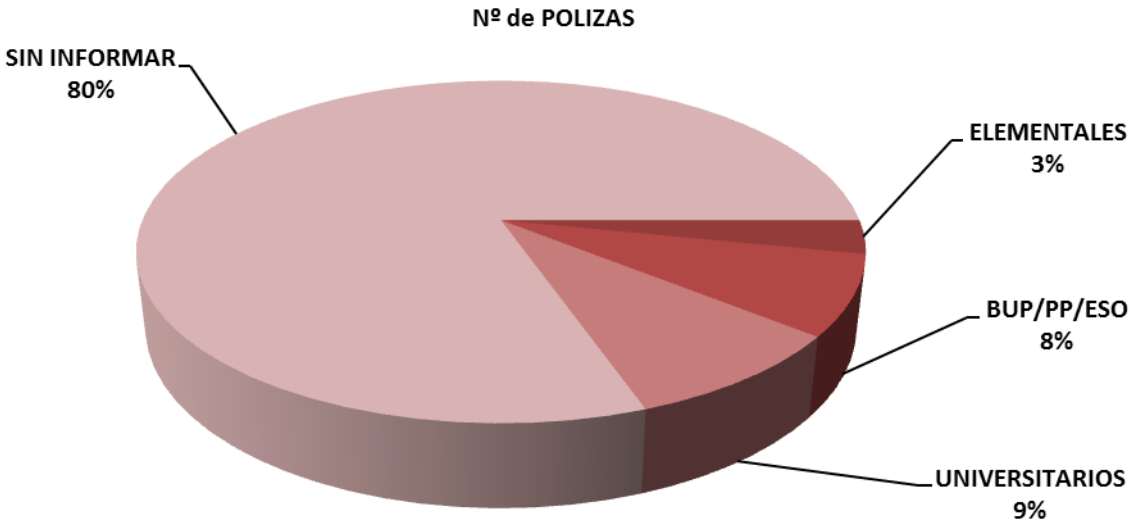


**Figura 23:** *Gráfico de la Distribución por NIVEL DE ESTUDIOS*
**Fuente**: *Propia de los Autores*

Así pues, al igual que sucede con la variable anterior, se observa que el COD=99 (Sin informar), es el más representativo por lo que también de ser considerado durante en el análisis de resultados, ya que se trata básicamente de una ausencia de información.

Nuevamente, la información del sector confirma que el nivel de estudios indica el nivel de aseguramiento tiende a ser creciente conforme se eleva el nivel educativo del responsable del hogar (*Tabla 24*):

Penetración de diferentes seguros en los hogares, según el nivel de estudios del sustentador principal.

| | AUTOS | SALUD | VIVIENDA | DECESOS | RC |
|---|---|---|---|---|---|
| Sin estudios o de primer grado | 62,1% | 20,1% | 62,7% | 67,5% | 11,1% |
| Educación secundaria, primer ciclo | 81,3% | 25,3% | 75,1% | 57,8% | 12,0% |
| Educación secundaria, segundo ciclo | 85,9% | 31,6% | 78,8% | 46,0% | 12,6% |
| Educación superior | 90,0% | 39,0% | 84,6% | 37,3% | 13,4% |

**Tabla 24:** *Tasas de Penetración según el nivel de estudios*
**Fuente**: *ICEA, Memoria Social del Seguro Español 2013*

### 3.5. Discretización de Variables

Previo a la aplicación, se debe mencionar la necesidad de realizar una *Transformación de Datos*, es decir, una adecuación de la base de datos mediante la conversión de los datos (Segovia-Vargas, 2003). Esto se debe a que comúnmente los datos pueden ser de tipo continuo o discreto.

Por un lado, en el caso de las variables cualitativas, se pueden trabajar como variables discretas de tal forma que se cuenta con un número finito de atributos que pueden tomar la variables En estos casos, se ha creado una serie de "códigos" que interpretan o clasifican el conjunto de variables ya descritas; es decir, traducir sus valores a términos absolutos utilizando valores números o a cierto carácter definido.

Siendo así, se realizaron las transformaciones mediante los códigos asignados a las variables de la siguiente forma:

- **SEXO**

| COD | SEXO |
|---|---|
| H | Hembra |
| V | Varón |

- **TIPO PRODUCTO**

| COD | TIPO_PRODUCTO |
|---|---|
| 1 | INDIVIDUAL AHORRO |
| 2 | INDIVIDUAL RIESGO |

- **TIPO PRIMA**

| COD | TIPO PRIMA |
|---|---|
| 1 | PRIMA UNICA |
| 2 | PRIMA PERIODICA |

- **RED**

| COD | RED |
|---|---|
| 1 | PROPIETARIAS (Agentes, Empleados) |
| 2 | NO PROPIETARIAS (Brokers, Corredores) |

- **FORMA PAGO**

| COD | FORMA PAGO |
|---|---|
| 1 | ANUAL |
| 2 | SEMESTRAL |
| 3 | TRIMESTRAL |
| 4 | BIMESTRAL |
| 5 | MENSUAL |
| 6 | UNICA |

- **ESTADO CIVIL**

| COD | EDO_CIVIL |
|---|---|
| 1 | Soltero |
| 2 | Casado |
| 3 | Divorciado |
| 4 | Viudo |
| 5 | Separado |
| 6 | Pareja de Hecho |
| 9 | Sin Informar |

- **HIJOS**

| COD | HIJOS |
|---|---|
| N | Sin hijos |
| S | Con hijos |

- **VALOR CLIENTE**

| COD | VALOR_CLIENTE |
|---|---|
| A | Muy Rentables y Muy Vinculados |
| B | Muy Rentables y Medianamente Vinculados |
| C | Muy Rentables y No Vinculados |
| D | Medianamente Rentables y Vinculados |
| E | Medianamente Rentables y No Vinculados |
| F | No Rentables y Vinculados |
| G | No Rentables y No Vinculados |

- **INDICE DE CAPACIDAD ECONOMICA**

| COD | ICE |
|---|---|
| 0 | Sin informar |
| 1 | Bajo |
| 2 | Medio-Bajo |
| 3 | Medio |
| 4 | Medio-Alto |
| 5 | Alto |

- **NIVEL INGRESOS**

| COD | NIV_INGRESOS |
|---|---|
| 1 | < 6.000 |
| 2 | 6.000 A 18.000 |
| 3 | 18.001 A 36.000 |
| 4 | 36.001 A 60.000 |
| 5 | 60.001 A 100.000 |
| 6 | 100.001 A 300.000 |
| 7 | 300.001 A 600.000 |
| 8 | > 600.001 |
| 9 | NO INFORMADO |

- **NIVEL ESTUDIOS**

| COD | NIV_ESTUDIOS |
|---|---|
| 01 | ELEMENTALES |
| 02 | BUP/PP/ESO |
| 03 | UNIVERSITARIOS |
| 99 | SIN INFORMAR |

Ahora bien, por otro lado, existen variables cuantitativas las cuales deben ser tratadas de manera diferente para poder ser traducidos en términos cualitativos. El empleo de este tipo de información implica una división del dominio original en algunos subintervalos; así como su correspondiente asignación de códigos cualitativos a dichos subintevalos (Segovia-Vargas, 2003).

Esta manipulación o discretización no viene impuesta por las técnicas de Inteligencia Artificial, sin embargo, la aplicación de la metodología y la posterior interpretación de los resultados finales es más sencilla. Por otro lado, no existe una única forma para establecer los subintervalos; por lo que se tomará la recomendación que se utiliza frecuentemente en los trabajos de investigación (Laitinen (1992), García

et al., (1997), McKee, (2000) o Segovia-Vargas (2003), que es el uso de percentiles que siguen las distribuciones en las variables continuas.

La única variable continua que se tiene en la muestra es la correspondiente a la EDAD. Siguiendo esta recomendación, se han calculado los percentiles: 20, 40, 60 y 80 para esta variable; quedando así su dominio dividido en cinco partes; asignando un código en orden ascendente dado que no existe algún criterio que haga pensar que a cierta edad mejora o empeora el subintervalo. Por lo que la discretización y asignación de códigos para esta variable quedaría de la siguiente forma:

- **EDAD (en intervalos)**

| COD | RANGO DE EDAD | PERCENTIL | OBSERVADO |
|---|---|---|---|
| 1 | [ 0 - 37 ) | | |
| 2 | [ 37 - 43 ) | P20 | 37 |
| 3 | [ 43 - 49 ) | P40 | 43 |
| 4 | [ 49 - 58 ) | P60 | 49 |
| 5 | [ 58 - y más ) | P80 | 58 |

De esta forma, se obtiene finalmente la base de datos *Transformada* (variables continuas discretizadas y las discretas con su valor original); de tal forma esta nueva tabla codificada ya pueda ser utilizada para la aplicación de las técnicas de Inteligencia Artificial.

# CAPITULO 4: APLICACIÓN DE LAS TÉCNICAS DE INTELIGENCIA ARTIFICIAL

## 4.1. Introducción

Como ya se ha mencionado, una de las principales aportaciones que exigen la nueva regulación de Solvencia II, es la gestión y control de los riesgos del sector asegurador. Por otro lado, la necesidad de medir el Riesgo de Caída de Cartera y promover que las entidades aseguradoras hagan una correcta evaluación de dicho riesgo, ha sido un esfuerzo del sector asegurador en su globalidad. Para conseguir dicha cuantificación y control sobre el Riesgo de Caída de Cartera, se puede lograr en la medida en la que se desarrollen modelos predictivos de cancelaciones potentes, unido a la identificación de las causas de anulaciones que fortalezcan a dichos métodos estadísticos.

Existen diversas metodologías utilizadas para la estimación de las anulaciones; sin embargo, la literatura actuarial sobre el tema de caída de cartera no ha sido muy extensa. Por el contrario, gran parte de las referencias que se tienen sobre el tema, se enfocan más hacia el estudio de la fidelidad de los asegurados. Es así como las primeras referencias encontradas se basan en el estudio de factores que incurren sobre la demanda de tipos de productos de seguros (Hammond et al. 1967). En la década de los ochenta, se encuentran los primeros trabajos sobre la retención y fidelidad de los asegurados: estudios de marketing relacional en la satisfacción, retención y precios en la industria del seguro de vida (Crosby y Stephens 1987). Poco después, se tiene la primera investigación donde se determina el valor del cliente (*Custome'sr Lifetime Value, CLV*) en el sector asegurador (Jackson 1989). Pasado los años, durante la década de los noventa, se retoma el tema de estudiar los factores que inducían a los clientes a cambiar de entidad para intentar aumentar la fidelidad en el seguro del automóvil (Schlesinger y Schulenburg 1993). En cuanto a estrategias de fidelización en el sector asegurador, se realizó un estudio en el ámbito de los seguros de salud basado en técnicas de segmentación (Cooley 2002).

Por lo que respecta al tema de cancelaciones, más tarde se analizaron los factores que incidían en la probabilidad de cancelación por parte de los clientes con varios contratos en la misma compañía (Brockett, et al. 2008). A partir de aquí, era posible establecer recomendaciones generales para gestionar el riesgo de negocio en las compañías aseguradoras (Guillén et al. 2008). Recientemente, la intensidad y consecuencias producidas por el riesgo de caída de cartera han sido descritas para el ramo de vida (Pieschacon 2010).

Es así como, con esta revisión de la literatura, se deja ver la necesidad de estudiar la caída de cartera, como el evento que se producirán en el futuro y haría fluctuar el volumen del negocio y márgenes de rentabilidad; que se traduce en la probabilidad de cancelación del contrato de seguros basado en la experiencia registrada en años anteriores. La mayoría de las ocasiones se recurre a técnicas estadísticas que, mediante un coeficiente de caída, recogen el promedio de porcentajes de caída registrados durante el histórico de la cartera.

Sin embargo, la utilización de dichas técnicas muestra poco margen de maniobra en cuanto a la gestión del riesgo como tal; ya que la visión puramente matemática que proporcionan estas metodologías, niegan la posibilidad de la inclusión de componentes cualitativos que maticen el resultado de tal forma que se pueda incurrir en él. En otras palabras, mediante una adecuada definición del "Apetito de Riesgo" o Nivel de Riesgo que pretenda una entidad aseguradora y el estudio de una serie de factores cualitativos que incurren en la decisión de permanencia o abandono en un cliente, se puede lograr una gestión y control del riesgo de caída de cartera mucho más manipulable y alineada con la estrategia de negocio planteada por la entidad.

Por otro lado, existen pocos estudios que analizan la estimación del riesgo de caída de cartera al que está expuesta una entidad aseguradora aplicando técnicas de Inteligencia Artificial (Martínez-Campos, 2014). Siendo así, el objetivo del presente capítulo, mediante la utilización de dichas metodologías no paramétricas que no requieren supuestos distribucionales, es detectar interacciones o relaciones no lineales para lograr identificar una serie de patrones de conducta que caracterizan a los

asegurados que buscan la anulación de su contrato de seguros. De otra forma dicho y basándonos en el principio básico de la Inteligencia Artificial, lograr establecer una serie de reglas de decisión básicas, a manera de herramienta de clasificación, que puedan ser capaces de determinar el perfiles clientes susceptibles a la cancelación de su póliza.

Con el objetivo de aportar un mecanismo de alarma o indicador de propensión al abandono de una póliza de Vida en una compañía aseguradora, se plantea una aplicación práctica basado en una muestra de pólizas de Seguros de Vida Individual. El seguro de vida es uno de los tipos de seguro en el que el pago de la suma asegurada del contrato por parte de la compañía de seguros depende del fallecimiento o supervivencia del asegurado en un momento determinado. En este tipo de seguro el pago de la indemnización no guarda relación con el valor del daño producido por la concurrencia del siniestro, debido a que la persona no es "valorable" económicamente. De ahí que este tipo de seguro no constituya un contrato de indemnización propiamente dicho, diferenciándose así, de los seguros de daños.

Para ello, abordaremos el tema primeramente con una revisión de las técnicas de Inteligencia Artificial utilizadas para la obtención de factores de comportamiento que definirían a los posibles clientes próximos a cancelar su contrato de seguros. Se revisarán las principales características de dos de las técnicas de la Inteligencia Artificial: Árboles de Decisión y Rough Set centrándonos principalmente en el algoritmo que demuestra su funcionamiento. Posteriormente, se realizarán una aplicación empírica de ambas técnicas sobre una cartera real de clientes de una compañía de seguros. Así se finalizará con una sección enfocada a los principales resultados obtenidos con el fin de identificar las características que puedan incurrir en el tipo de cliente susceptible a la anulación de su póliza, mediante los resultados obtenidos de ambas técnicas.

En el contexto actual del mercado asegurador en donde existe una disminución del volumen de negocio y tendencia creciente a la pérdida de la cartera de clientes; cobra importancia el tema de retención de clientes y con ello suena interesante poder identificar el tipo de clientes propensos a causar baja. De esta forma, se podrán

anticipar pérdidas mediante la implementación de estrategias para la retención y atracción de nuevos clientes; es decir, lograr orientar la toma de decisiones por medio de la localización de algún patrón de comportamiento del tipo de cliente “cancelador” que permita establecer políticas comerciales atractivas para la captación y fidelización de su cartera. De aquí la relevancia de la presente aplicación empírica, para poder tener una aproximación a la probabilidad de cancelación del cliente mediante dichos patrones que se traduciría en una mejora en la gestión del riesgo de caída de cartera que a su vez, contribuiría al equilibrio y estabilidad de los niveles de solvencia que las compañías aseguradoras requieren.

### 4.2. Inteligencia Artificial

En esta sección, se describe la metodología utilizada en la aplicación empírica que se ha realizado posteriormente. Se trata de una metodología compleja en cuanto a su configuración pero se torna sencilla en cuanto a la interpretación de los resultados obtenidos, así como el análisis e implicaciones que se obtienen a partir dichos resultados. Una de las definiciones hecha por uno de los pioneros de esta metodología quien dice que "la Inteligencia Artificial es la ciencia de construir máquinas para que hagan cosas que, sí las hicieran los humanos, requerirían inteligencia" (Minsky Marvin, 1967). Así pues, se abordará uno de los campos que engloba las técnicas de Inteligencia Artificial, lo cual estudia la creación y diseño de sistemas capaces de resolver cuestiones por sí mismas utilizando como modelo la propia inteligencia humana (Galipienso, María Isabel Alfonso, et. al, 2003).

En primer lugar, se debe situar a la "Inteligencia Artificial" dentro de las metodologías que se manejan en la rama de la disciplina de *Aprendizaje Automático (Machine Learning,* por sus siglas en inglés). Este enfoque utiliza algoritmos para analizar registros en bases de datos internas de los clientes de una empresa, para descubrir ciertos patrones, interacciones o reglas que pueden describir o predecir las futuras tendencias que puedan indicar cualquier tipo de oportunidades competitiva (Mena, 1996), ayude a tomar decisiones o mejorar la comprensión o conocimiento que se pueda extraer a través de dichas bases de datos. Es decir, se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones en los mismos que puedan ser usados para realizar predicciones válidas; unido a ello, se puede decir que otra de sus ventajas es el bajo coste computacional que ofrece este tipo de algoritmos.

Ahora bien, dentro de esta disciplina se engloban las técnicas de Inteligencia Artificial, las cuales se basan en el aprendizaje a partir de los datos y de su semejanza con un pensamiento estructurado similar al comportamiento humano. Siendo tan amplio su campo de acción, reúne varias áreas de investigación donde ha sido utilizada; donde uno de éstos es el reconocimiento de patrones con el propósito de

extraer información que permita establecer propiedades y características de cierto conjunto de objetos.

Existen varias técnicas sugeridas dentro de esta rama de la Inteligencia Artificial como son las Redes Neuronales, los Vectores Soporte, los Algoritmos Genéticos, los Sistemas de Inducción de Reglas, los árboles de Decisión o la Teoría de Rough Set.

Algunos autores han validado estas técnicas mediante su aplicación a diversos datos reales para ciertas líneas de investigación y estudios realizados. Por mencionar algunos de ellos, cabe mencionar Sanchis, et al. (2007) y Díaz et al. (2009) enfocándose en la parte de Reglas y Árboles de Decisión, respectivamente. En lo que se refiere a la Teoría Rough Set se debe mencionar a Segovia-Vargas (2005). Centrándose las aplicaciones empíricas no paramétricas del presente trabajo en estas dos últimas técnicas, a continuación se detallan recurriendo en gran parte a los estudios realizados por dichos autores.

### 4.2.1. Técnica de Arboles de Decisión

El concepto de árbol de decisión se engloba dentro del ámbito de la Inteligencia Artificial como uno de los modelos predictivos que estudia. Ahora bien, a partir de un conjunto de datos se construyen diagramas de construcciones lógicas que hacen referencia a una clasificación óptima de los datos de acuerdo a sus características o atributos. De esta forma, se crean particiones recursivas que sirven para categorizar y representar una serie de condiciones que ocurren de forma sucesivas, comúnmente llamadas reglas sobre la decisión que se debe tomar, para solucionar el problema planteado asignando un valor de salida a un determinado registro de entrada. Dichas reglas, gráficamente, se representan en forma de árbol a través de hojas o ramas; de ahí el nombre de Árbol de Decisión; de esta forma, permite obtener de forma visual, las reglas de decisión, de aquí su principal ventaja que es la fácil interpretación de los resultados.

Existen varios algoritmos para la construcción de árboles de decisión: CLS (*Concept Learning Systems*; Hoveland y Hunt, 1950), Método CHAID (G.V. Kass, 1980), Método CART (Breiman, Friendman, Olshen y Stone, 1984), Algoritmo C4.5 (J.R. Quinlan ,1994). La diferencia entre estos algoritmos de aprendizaje radica en el criterio utilizado para realizar las particiones o *Reglas*; esto en otras palabras habla de la definición de la partición óptima de un nodo. Esta es la ventaja de los algoritmos llamados Método ID3 (Quinlan J.R., 1973 y 1986) y sus sucesores como C4.5, así como otras mejoras de éste como es el C5.0. Es así, como el Algoritmo C4.5 es uno de los algoritmos más utilizados en el ámbito de los árboles de Decisión; y por ende, la razón por la cual fue elegido este tipo de algoritmo para su uso en la aplicación empírica realizada.

El algoritmo C4.5 se basa en conceptos procedentes de la Teoría de la Información para hacer las particiones y fue desarrollado por Quinlan (Quinlan, J. R., 2014). Para detallar brevemente los conceptos esenciales de este tipo de algoritmo, se utilizará el trabajo de autores expertos en el tema (Díaz, et. al 2009); ya que han aplicado este tipo de técnicas en muestras de datos españoles.

El C4.5, parte de la premisa de tomar en cada rama del árbol, para hacer la correspondiente partición, aquella variable que proporciona más información de cara a clasificar los elementos que constituyen el conjunto de entrenamiento o conjunto de datos usados para construir el árbol. Para establecer la variable que proporciona la mayor información, en el caso del C4.5 se emplea el *ratio de ganancia (Gain Ratio*).

Se puede observar, mediante el ejemplo (*Figura 24*), que la interpretación de los resultados es sencilla y es fácil seguir la lógica que se debe seguir para su aplicación a través del recorrido de sus Reglas o ramas del árbol dibujado. De ahí su atractivo ya que puede ser analizado incluso por personas con poca experiencia en el tema.

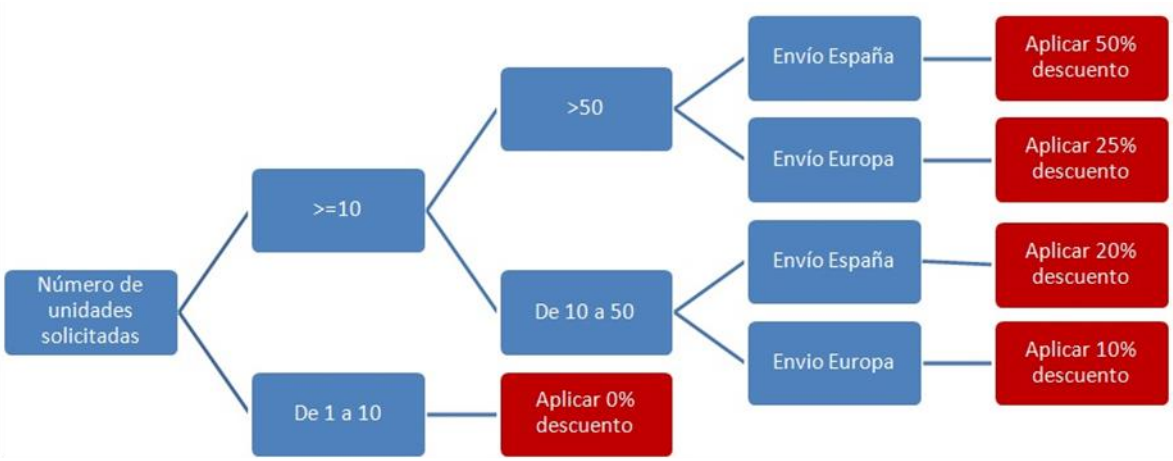


**Figura 24:** *Ejemplo de Árbol. De Decisión*
**Fuente**: *Propia de los autores*

En cuanto a su estructura, se distinguen los siguientes componentes:

- *Nodo Interno*: Consiste en una pregunta o test relativa al valor de un atributo. De cada nodo interno parten tantas ramas como respuestas haya a la pregunta, que normalmente equivale al número de posibles valores que puede tener el atributo en cuestión

- *Nodo Hoja*: En cada nodo hoja sólo puede haber instancias (casos) con un único valor de clase

- *Ramas*: Son las divisiones excluyentes y exhaustivas del conjunto de elementos que se quieren clasificar

Ahora bien, para construir el árbol de decisión se utiliza la estrategia de "divide y vencerás"; esto es, a través de un algoritmo se realizan divisiones sucesivas del espacio multivariable para maximizar la distancia entre grupos en cada división. Este proceso de división finaliza cuando todos los registro de una rama tienen el mismo valor en la variable de salida dando lugar al modelo completo.

A la variable de salida también se le conoce como *Nodo Hoja Puro*; que es aquel al que sólo corresponden casos pertenecientes a una de las clases del problema, o cuando la ramificación del árbol ya no suponga ninguna mejora.

Se entiende que cuanto más abajo están las variables de entrada en el árbol, quiere decir que menos generalización permite, y por tanto, menos importantes son en la clasificación de salida.

Ahora bien, para dejar resumido con mayor claridad el proceso que se sigue, se retoman cuatro etapas definidas por los autores (Esquerda, Aureli, et al. 2007):

i. Desarrollo del árbol

A partir del *Nodo Raíz*, se identifica la variable más adecuada para dividir dicho nodo en dos *Nodos Hijo*. A cada uno de estos nodos, se les asigna un valor de la variable dependiente que se corresponde al mayor número de registros de ese nodo. Y a su vez, cada *Nodo Hijo* será subdividido en nuevos nodos sucesivos para seguir el proceso

ii. Parada del desarrollo

Esta etapa hace referencia al momento en el que desarrollo del árbol se detiene. Esto sucede cuando los *Nodos Hijos* ya no pueden subdividirse ya que contienen un único caso; o bien cuando el valor de la variable dependiente es el mismo para todos los casos integrantes del nodo.

iii. Poda del árbol

Aquí se trata de eliminar las ramas con pocos registros o poco significativas; es decir, se eliminan o podan las condiciones de las ramas del árbol de tal forma, que se obtengan modelos más generales con mayor error de clasificación sobre el conjunto de casos de entrenamiento pero menor sobre nuevos casos no usados en la construcción del árbol. En otras palabras, en la fase en la que elimina aquello cuya presencia añade más complejidad que efectividad.

iv. Selección del árbol óptimo

Hace referencia a la fase en la que se elige el árbol óptimo que mejor clasifica al grupo de validación. Para ello, se necesita de un sistema de validación; la cual puede ser externa, utilizando casos no empleados en el desarrollo del modelo; o bien, interna o validación cruzada. Esta última se trata de realizar una partición aleatoria del grupo de desarrollo; primero se utiliza de forma recursiva en un subgrupo para generar el árbol y se valida en un segundo subgrupo. Ahora bien, el objetivo de este proceso es la obtención de un árbol lo más simple y predictivo posible, y que garantice una solución óptima; de aquí la existencia de varios algoritmos para la construcción de los árboles.

La información que proporciona un mensaje o la realización de una variable aleatoria x es inversamente proporcional a su probabilidad $\mathrm{P_x}$ (Reza, F. M., 1961). Con frecuencia en Ingeniería de Comunicaciones o en Estadística se mide esta cantidad en bits, que se obtienen como $\log_2 \frac{1}{\mathrm{P_x}}$. El promedio de esta magnitud para todas las posibles ocurrencias de la variable aleatoria $\mathrm{x}$ recibe el nombre de entropía de $\mathrm{x}$, es decir, el promedio se obtendría multiplicando los posibles estados que puede tomar la variable $\mathrm{x}$, $\log_2 \frac{1}{\mathrm{P_x}}$, por su probabilidad de ocurrencia, $\mathrm{p(x)}$. Luego la entropía de $\mathrm{x}$ será, $\mathrm{H(x)}$:

$$H(x) = \sum_x p(x)\, log_2 \frac{1}{p_x}$$

En consecuencia, la entropía es una medida de la aleatoriedad o incertidumbre de $\mathrm{x}$ o de la cantidad de información que, en promedio, nos proporciona conocimiento de $\mathrm{x}$.

De manera similar se define la entropía conjunta $\mathrm{H(x, y)}$, para ello se parte de dos variables aleatorias $\mathrm{x}$ e $\mathrm{y}$:

$$H(x,y) = \sum_{x,y} p(x,y)\, log_2 \frac{1}{p_{(x,y)}}$$

La entropía conjunta es la cantidad de información que, en promedio, nos proporciona el conocimiento de $\mathrm{x}$ e $\mathrm{y}$.

A partir de los conceptos vistos se puede definir otros relacionados como la entropía condicional de x dada y, $H(x\backslash y)$, se define como:

$$H(x\backslash y) = \sum_{x,y} p(x,y)\, log_2 \frac{1}{p_{(x\backslash y)}}$$

La entropía condicional es una medida de la incertidumbre respecto a x cuando se conoce y. Representa la cantidad de información que se necesita para conocer plenamente x cuando ya se tiene la información suministrada por y.

Obviamente se cumple que $\mathrm{H(x\backslash y)} \leq \mathrm{H(x)}$, pues al conocer y se tiene más información que pueda ayudar a reducir la incertidumbre sobre x.

Esto permite reducir la incertidumbre y a la misma se la denomina información mutua entre x e y: $\mathrm{I(x;y)} = \mathrm{H(x)} - \mathrm{H(x\backslash y)}$, ya que es la información que una de las variables se transmite sobre la otra. Además se verifica que $\mathrm{I(x;y)} = \mathrm{I(y;x)}$, siendo la información mutua una magnitud similar a la covarianza.

Originariamente (Quinlan, 1973 y 1986), se seleccionaba para hacer cada partición aquella variable y que proporcionaba la máxima información sobre x, es decir, maximizaba $\mathrm{I(x;y_i)}$ (magnitud denominada Gain). Sin embargo, maximizar *"gain"* proporciona buenos resultados, pero introduce un sesgo en favor de las $\mathrm{y_i}$ con muchos valores distintos.

En las versiones posteriores del algoritmo se corrige este sesgo y se selecciona aquella $\mathrm{y_i}$ que maximiza la magnitud $\frac{\mathrm{I(x;y_i)}}{\mathrm{H(y_i)}}$ (denomina Gain Ratio). Se define como el porcentaje de la información proporcionada por $\mathrm{y_i}$ que es útil para conocer x.

Podría ocurrir que un atributo se seleccionara básicamente porque su entropía $\mathrm{H(y_i)}$, es pequeña, lo que aumentaría el valor del cociente anterior. Para evitarlo se exige además que $\mathrm{I(x;y_i)}$ sea razonablemente grande.

Si el procedimiento descrito, se aplica de manera reiterada se va construyendo el árbol de decisión; hasta que se alcanza la pureza del nodo y con ello finaliza el proceso.

En el algoritmo C4.5, para facilitar la comprensión del árbol, se puede realizar una *poda* del mismo. El proceso de *poda* comienza en los *Nodos Hoja* y recursivamente continúa hasta llegar al *Nodo Raíz*. En consecuencia, tras la *poda* del árbol, éste ganará en capacidad de generalización, a costa de reducir el grado de pureza de sus hojas (Hernández et al., 2004 y Larose, 2005). Es decir, se obtienen modelos más generales pero aumentando el error de clasificación.

Ahora bien, los conceptos explicados se pueden concretar de la siguiente forma para un problema cualquiera:

- Se puede considerar que es una variable aleatoria que muestra la clase a la que pertenece un elemento,

- Y se considera a $y_i$, siendo $i = 1,2,\dots,n$, son los atributos o variables que caracterizan a los elementos que se quieren clasificar.

De esta forma, para aplicar el algoritmo C4.5 a la cartera-muestra utilizada para la aplicación empírica y análisis de la caída de cartera, se traduce en que:

- La variable $x$ indicará si la póliza está en vigor o anulada,

- Las variables $y_i$ serán las 14 variables cualitativas utilizadas para la clasificación

A través de los valores que vayan tomando estas variables cualitativas, se irá *condicionando y reduciendo* la incertidumbre de $x$, vigor o anulada, e irá dependiendo de la información que vayan proporcionando las 14 variables.

Ahora bien, para aplicar el Algoritmo C4.5 se tienen varias posibilidades. Se puede descargar gratuitamente la versión C4.5 Release 8 desde la página de Ross Quinlan[31]. Este programa, una vez compilado, puede ser ejecutado en sistemas operativos *Unix*. Además, existen nuevas versiones comerciales del algoritmo (C5.0 para Unix y See5 para Windows) que implementan mejoras y funcionalidades adicionales y se comercializan por su creador, Quinlan, (RULEQUEST RESEARCH) o a

[31] http://www.rulequest.com/Personal/

través de paquetes de minería de datos como Clementine, aunque también hay versiones de demostración gratuitas limitadas a bases de datos pequeñas[32].

Así mismo, en cuanto al sistema de inducción de árboles de decisión del C5.0, parece ser esencialmente el mismo que en C4.5. Sin embargo, la inducción de reglas con las nuevas versiones es diferente y más rápida. En nuestro estudio no induciremos reglas de decisión (donde sustancialmente están las mejoras del C5.0) por lo que la utilización del C4.5 es suficiente

Concretamente, como ya se ha comentado, se ha decidido realizar la aplicación empírica mediante el algoritmo J48, la cual es la implementación en Java de libre acceso del algoritmo C4.5 y que contiene la herramienta de minería de datos WEKA, el cual es el paquete de minería de datos desarrollado por la Universidad de Waikato (Witten, Ian H., et al., 1999).

### 4.2.2. Teoría de Rough Set

Como ya se ha puesto de manifiesto, la Inteligencia Artificial cuenta con numerosas técnicas dentro de las cuales se tienen los algoritmos de Inducción de Reglas. Es aquí donde se engloba la Teoría de Rough Sets (*Método de Conjuntos Aproximados*); la cual ha demostrado una gran eficacia cuando existe un conjunto de datos caracterizado por la misma información pero clasificados en grupos distintos; lo cual es común cuando se trabaja con bases de datos reales, como es el caso que se tiene en cuestión.

De igual forma que se hiciese con la técnica anteriormente descrita, se recurrirá a diversos autores quienes han podido aplicar esta metodología sobre diversas problemáticas planteadas para dar reseña y repaso de los detalles y ventajas de esta segunda metodología propuesta para el estudio empírico por desarrollar.

[32] http://www.rulequest.com/

Primeramente, cabe mencionar que esta teoría fue introducida en el año 1982 por Pawlak como una nueva técnica de gran utilidad para el análisis y contenido de tablas de información que describen a un conjunto de objetivos por medio de una serie de atributos. Aunque hoy en día existen extensiones de esta metodología (Greco, S., Matarazzo, B. y Slowinski, R., 1998), se expondrá el enfoque clásico, mismo que ha sido el utilizado en la aplicación posterior.

Esta teoría utiliza la experiencia en eventos pasados acumulados sobre una serie de patrones de datos, para finalmente poder obtener una serie de reglas en forma de sentencias lógicas que nos ayuden en la toma de decisiones futuras. Para ello, el enfoque de Rough Set esta fundamentalmente basado en un proceso de toma de decisiones. Es así como hace necesario referenciar que un problema de decisión implica un conjunto de *objetos* descritos por un conjunto de *atributos*; el cual se puede representar mediante una *tabla de decisión;* así se tiene que uno o varios agentes (expertos, decisores, accionistas, etc.) están implicados en el problema de decisión.

Ahora bien, sabiendo que se parte de una tabla de decisión que incluye información sobre la posible toma de decisiones, ésta puede incluir cierto tipo de preferencias del agente que apoyen las nuevas decisiones por considerar, lo cual hace referencia a un *modelo global sobre las preferencias* (Roubens, M. y Vincke, P., 2012). De aquí, hay dos formas para construir este tipo de modelos: modelo funcional o modelo relacional. Este segundo es donde se engloba los Rough Set, según Segovia (2003).

El modelo relacional se basa en el aprendizaje de los ejemplos o adquisición del conocimiento inductiva (inducción de reglas, aprendizaje inductivo) (Michalski, R. S. 1983). Este enfoque ofrece más confianza a la valoración efectivamente realizada por un agente que la explicación que tuviera que dar sobre la misma dicho agente. El modelo resultante es un conjunto de reglas de la forma Si / Entonces o bien un árbol de decisión; las cuales son fáciles de comprender por los usuarios finales.

Pues bien, retomando el marco de la Inteligencia Artificial, la información de las *preferencias* se denomina conocimiento acerca de las preferencias y al decisor se le denomina *experto*. Es así como el enfoque Rough Set se engloba dentro de este tipo de

modelos de preferencia global cuyo objetivo final es la obtención de reglas de decisión. La cuestión consiste en determinar una serie de reglas que nos ayuden a determinar si cada uno de los objetos del sistema pertenece al conjunto denominado clase de decisión. Dicha regla de decisión puede estar representada como una sentencia lógica con la siguiente forma: Si (se cumple la condición) entonces (pertenece a la clase).

Sin embargo, se debe hablar sobre las inconsistencias de los objetos que puede llevar a la ambigüedad en su clasificación; es decir objetos descritos por los mismos atributos pero asignados a diferentes clases (Roy, R. y Kailath, T. 1989). Esto es lo que diversos sistemas definen y se ven obligados a trabajar con la presencia de "ruido". De aquí que la Teoría Rough Set sea útil cuando las clases en las que se han de clasificar los objetivos son imprecisas, pudiendo aproximarse hacia conjuntos precisos (Nurmi et al. 1996).

De esta forma, se puede enumerar las siguientes ventajas que caracterizan a la Teoría Rough Set destacando:

- Utilización de variables tanto de tipo cuantitativo como cualitativo
- No necesita de ningún tipo de información preliminar o adicional de los datos como distribuciones de probabilidades estadísticas.
- Eliminación de variables redundantes y de esto modo enfocarnos en conjuntos mínimos de variables logrando una reducción del costo y tiempo del proceso asumido por el centro decisor
- Obtención de una serie de reglas de decisión de fácil comprensión. Así mismo dichas reglas están bien respaldadas por experiencia pasada lo cual argumenta las decisiones que se toman.

Es así como la filosofía de la Teoría Rough Sets está basada en el supuesto de que cada uno de los objetos considerados en el universo en estudio se le puede asociar alguna información, de acuerdo con Segovia-Vargas (2003). De tal forma, los objetos caracterizados por la misma información no son discernibles a la a vista de la información disponible sobre ellos.

De aquí surge la relación de ―no diferenciación‖ de los objetos y que se traduce en una de las principales ventajas de este método en el análisis de datos, esto es, que puede trabajar con conjuntos de datos inciertos e imprecisos, pero sin embargo pueden aproximarse mediante conjuntos precisos.

La aproximación del espacio y la aproximación de un conjunto en este espacio son dos conceptos de gran importancia en la Teoría de Rough Set, ya que en otras palabras, un "rough set" es una colección de objetos no clasificados de forma precisa en términos de los atributos, mientras que las aproximaciones que realiza por arriba y por debajo si lo hacen. Esto da como resultado casos fronterizos de objetos que no pueden ser clasificados con certeza. A partir de aquí se puede definir la *precisión* de la aproximación y la *calidad* de la misma (intervalos entre 0 y 1); mediante la *aproximación por encima* que contiene los objetos que posiblemente pertenecen al conjunto; y la *aproximación por debajo* con todos los objetos que con seguridad pertenecen al conjunto.

Otra de las características de la Teoría Rough Set es la capacidad de clasificación, con lo cual se pueden formar clases de objetos de acuerdo a las diferencias que presentan entre ellos. Esto es lo que, de acuerdo a Segovia-Vargas (2003), se enmarca como conocimiento y es de gran importancia para definir los conceptos claves de esta Teoría: *aproximación, dependencia y reductos*.

La Teoría Rough Set representa dicho conocimiento de los objetos en forma de una tabla de información. De esta forma, en la filas x se indican los objetos (acciones, empresas, etc.) y en las columnas q se representan los atributos. Así, los valores del atributo son las entradas de la tabla tomando el valor f(x,q). De esta forma, cada fila en la tabla representa la información sobre un objeto S; siendo éste el sistema de información denominado como sistema de representación del conocimiento. Además, el conjunto de atributos se divide en un subconjunto de atributos de condición y otro subconjunto de atributos de decisión; siendo así, si se distingue entre ambos conjuntos, se obtiene la tabla de decisión.

A manera de ejemplo, se ilustra *(Tabla 25)*, (como lo hiciese Moscarola, 1978 y Slowinski, 1993) para clarificar los conceptos anteriores:

| Criterios / Candidatos | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | D |
|---|---|---|---|---|---|---|---|---|
| $X_1$ | 4 | 4 | 4 | 4 | 2 | 2 | 1 | A |
| $X_2$ | 3 | 3 | 4 | 3 | 2 | 1 | 1 | R |
| $X_3$ | 3 | 4 | 3 | 3 | 1 | 2 | 2 | R |
| $X_4$ | 5 | 3 | 5 | 4 | 2 | 1 | 2 | A |
| $X_5$ | 4 | 4 | 5 | 4 | 2 | 2 | 1 | A |
| $X_6$ | 3 | 4 | 3 | 3 | 2 | 1 | 3 | R |
| $X_7$ | 4 | 4 | 5 | 4 | 2 | 2 | 2 | A |
| $X_8$ | 4 | 4 | 4 | 4 | 2 | 2 | 2 | A |
| $X_9$ | 4 | 4 | 4 | 4 | 2 | 2 | 2 | R |
| $X_{10}$ | 5 | 3 | 5 | 4 | 2 | 1 | 2 | A |
| $X_{11}$ | 5 | 4 | 4 | 4 | 1 | 1 | 2 | A |
| $X_{12}$ | 5 | 3 | 4 | 4 | 2 | 2 | 2 | A |
| $X_{13}$ | 4 | 3 | 3 | 3 | 3 | 2 | 2 | R |
| $X_{14}$ | 3 | 3 | 4 | 3 | 2 | 3 | 3 | R |
| $X_{15}$ | 4 | 5 | 5 | 4 | 2 | 1 | 1 | A |

**Tabla 25:** *Tabla de Decisión - Ejemplo*
**Fuente**: *Moscarola, 1978 y Slowinski, 1993*

Ahora bien, otro de los conceptos que surgen dentro de la Teoría Rough Set es la *dependencia causa-efecto* entre estos dos grupos de atributos, esto es, encontrar las relaciones entre todos los atributos y casos superfluos en el sistema de información, siendo lo más interesante del análisis de los sistemas de información. De esta forma, descubriendo dichas dependencias entre los atributos, se pretende lograr la reducción del conocimiento; esto es el proceso de expresar un conocimiento dado del modo más eficiente (Segovia-Vargas, 2003); mediante la *reducción* de todos los atributos.

Dicha *reducción* de atributos se consigue mediante la obtención de un modelo tal que el conjunto reducido de atributos proporcione la misma calidad de clasificación que el conjunto original de atributos; también denominado *conjunto mínimo de atributos o reducto*.

Ahora bien, de aquí surge otro término utilizado dentro de esta metodología, el *núcleo*. Esto es, la colección de atributos más relevantes en la tabla que no pueden ser eliminados sin que disminuya la calidad de aproximación dela clasificación. En otras palabras, el *núcleo* se compone de aquellas clasificaciones que son las más esenciales en el conocimiento; no pudiendo eliminar relación alguna del *núcleo* sin distorsionar el conocimiento. Y por el contrario, un *reducto* proporciona un conjunto de relaciones suficiente para caracterizar el conocimiento sin pérdida de información esencial. Es decir, en los *reductos* puede considerarse uno u otro; lo cual no sucede para el *núcleo* ya que es único al estar formado por las intersecciones de todos los reductos.

Así pues, Skowron propuso uno de los modelos más utilizados para representar el conocimiento en forma de una *matriz de diferenciación* (Skowron y Grzymala-Busse, 1991), misma que es simétrica y lo cual simplifica y facilita el cálculo del *núcleo* y *reductos* de una forma simple. Siguiendo el ejemplo propuesto por Pawlak (1982) se puede dejar claro este concepto. Observando en la matriz resultante (*Tabla 26*), se deduce que el núcleo es el atributo b y que existen dos reductos: el reducto {a,b} y el formado por {d,b}.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | a, b, c, d | | | | |
| 3 | a, b, c, | b, c, d | | | |
| 4 | a, c, d | a, b, d | a, b, c, d | | |
| 5 | a, c, d | b | b, c, d | a, d | |

**Tabla 26:** *Matriz de Diferenciación - Ejemplo*
**Fuente**: *Pawlak, 1982*

Ahora bien, otro de los conceptos tratados en este tipo de metodología son las *reglas de decisión*; siendo éstas el conjunto de datos que representan la experiencia. Se entiende que el conjunto de datos contiene información de un conjunto de *objetos* descritos por un conjunto de *atributos*. Por lo tanto, el tema consiste en encontrar reglas que determinen si un objeto pertenece a un subconjunto particular denominado *clase de decisión* o a un *concepto*. Como ya se ha mencionado, este tipo de reglas se

presentan en forma de sentencias lógicas: *SI <condiciones> ENTONCES <clases de decisión>.* Para formalizar el tema, Segovia-Vargas (2003) ha hecho una profundización en este sentido.

Así se tiene que el conjunto de reglas para todas las clases de decisión se denomina *Algoritmo de Decisión*; el cual puede entenderse como la representación más compacta; esto es, el menor número de reglas de decisión; y sin redundancias correspondiente a un sistema de información; lo cual se obtiene al tener el menor número de atributos que aparezcan en la definición de todas las reglas. Por lo tanto, esto hace que a su vez, el algoritmo de decisión sea más legible para el usuario que el sistema completo de información.

Ahora bien, cada regla de decisión se caracteriza por el número de objetos que satisfacen la parte de la condición de la regla y pertenecen a la clase de la decisión sugerida; esto es, lo denominado *fuerza* de la regla. Así se tiene que no todas las reglas son igual de importantes o fiables para el agente decisor: en cuanto *más débil* la regla es, *menos fiable* es en la toma de decisión.

Pueden establecerse dos perspectivas principales para la inducción de reglas de decisión derivadas de un conjunto, siendo las más comunes:

- *Inducción orientada a la clasificación*: cuyo objetivo es encontrar de forma automática, un conjunto de reglas que serán utilizadas para construir una clasificación de un conjunto de objetos.
- *Inducción orientada al descubrimiento*: cuyo objetivo es extraer patrones de información y regularidades "interesantes" y "útiles" para el usuario (dependiendo de sus exigencias y expectativas) que definan al mismo conjunto de objetos.

Ahora bien, para la aplicación empírica del presente trabajo, se utilizará el software informático RSES2[33], el cual se basa en este tipo de algoritmos que inducen reglas orientadas al descubrimiento; razón por lo que no se ha profundizado en todos los algoritmos desarrollados en la inducción de dichas reglas de decisión basados en el

[33] http://logic.mimuw.edu.pl/~rses/

enfoque de Rough Set. Sin embargo, cabe mencionar que el sistema que utiliza para obtener reglas que sean fuertes, simples y consistentes se utilizan y definen ciertos niveles *de fuerza, longitud* y *grado de discriminación*.

Es así como, finalmente el tema radica en la elección de las reglas de decisión óptimas que mejor describen al sistema de información. Skowron (1993) propuso un método para la generación de reglas de decisión óptimas con coeficientes ciertos; el cual se basa en la construcción de funciones booleanas apropiadas derivadas de las matrices de diferenciación modificadas. Por otro lado, el método también se puede aplicar mediante la construcción de reglas basadas en las aproximaciones por arriba y por debajo.

Finalmente, una vez que se obtiene el algoritmo de las reglas de decisión, el cual representa el conocimiento se obtuvo sobre los casos dentro de un sistema de información, sería interesante y deseable utilizar este conocimiento para justificar la clasificación de nuevos objetos; es decir, aquellos que no estén contemplados en el sistema de información inicial. Esto es, encontrar en el algoritmo de decisión aquella regla (o reglas cercanas), cuya parte de la condición coincida con la descripción del nuevo objeto.

Dado que para efectos de la aplicación presente, se trata de un tema de clasificación del tipo de clientes susceptibles a cancelar su póliza de seguros, no profundizaremos en otro tipo de problemas de decisión donde se aplica este tipo de metodología Rough Set. Sin embargo, de manera general, se pueden mencionar otros tres tipos de problemas de decisión:

a) Problemas de Clasificación con atributos múltiples: Consiste en asignar cada objeto a una categoría apropiada previamente definida, es decir donde sólo existe un único atributo de decisión
b) Problemas de Clasificación múltiple con atributos múltiples: En este caso, a diferencia del anterior, sí existen múltiples atributos de decisión, por ejemplo, el conjunto de casos que hay que clasificar proviene de varios agentes

c) Descripción de objeticos con atributos múltiples: Cuando los problemas están asociados con la explicación de una situación de decisión.

En el último caso c) es donde la Técnica de Rough set se adapta ya que el principal interés es buscar la mínima descripción posible en términos de atributos. Una descripción mínima permite un minucioso análisis de los conflictos, cuestión que resulta interesante en el momento de explicar o interpretar los resultados. Además si los atributos son consecuencias de algunas decisiones; lo que en ciertas aplicaciones se puede interpretar como relaciones “causa-efecto”; la metodología Rough Set permite descubrir las mínimas dependencias elementales entre las consecuencias. Lo cual traducido al contexto del presente trabajo sería de gran utilidad, ya que mediante mínimas interdependencias de las variables cualitativas que caracterizan a cada uno de los clientes dentro de una entidad aseguradora, se pueda definir el perfil de dichos clientes con mayor propensión a la anulación de su contrato de seguros.

### 4.3. Aplicación Empírica de la Técnica de Arboles de Decisión

En esta sección, se analizan los resultados obtenidos de aplicar un Árbol C4.5 a la cartera muestra. Este tipo de algoritmo se utiliza para analizar problemas de descripción y clasificación de objetos descritos por múltiples variables y asignados en una categoría determinada. En la aplicación del presente estudio, se tiene una cartera de pólizas caracterizadas por una serie de valores cualitativos y cuantitativos, que se intentan asignar a alguna de las dos categorías: vigor o anulada.

En la muestra que se tiene, la distinción de estas 2 categorías viene dado según el criterio que toma la variable de decisión *TIPO DE PRESTACION*. Esta variable toma el valor de 0 indicando que la póliza se encuentra en Vigor. Por el contrario, toma el valor 1 si la póliza se encuentra Anulada.

Ahora bien, para conocer la magnitud de la muestra, se ha obtenido el desglose de la muestra según la variable de decisión:

| COD | TIPO_PRESTACION | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 0 | VIGOR | 16,568 | *83.74%* |
| 1 | ANULADA | 3,216 | *16.26%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 27:** *Distribución de acuerdo a la variable TIPO DE PRESTACION*
**Fuente**: *Propia de los autores*

Ambas categorías han sido evaluadas de acuerdo a los valores que tomas las 14 variables cualitativas y cuantitativas seleccionadas *(Tabla 27)* considerando una misma base de datos. Esta base de datos se ha introducido y programado en *WEKA*, programa informático que ha desarrollado el análisis del algoritmo C4.5.

### 4.3.1. Resumen de Validación de Resultados bajo el Algoritmo C4.5

Primeramente, la salida bruta de los resultados obtenidos se muestra a continuación (*Figura 25*). Se puede observar que los resultados cuenta con un porcentaje de aciertos del 86.31% (*Correctly Classified Instances*) de acuerdo al Resumen de Validación de Resultados que arroja el programa WEKA; lo cual justifica su interpretación.

```
Number of Leaves:          445

Size of the tree:          633


Time taken to build model: 0.58 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       17076               86.3122 %
Incorrectly Classified Instances      2708               13.6878 %
Kappa statistic                          0.4111
Mean absolute error                      0.1868
Root mean squared error                  0.3221
Relative absolute error                 68.5924 %
Root relative squared error             87.2969 %
Total Number of Instances            19784

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.954     0.603     0.891     0.954     0.921     0.818     0
                 0.397     0.046     0.624     0.397     0.486     0.818     1
Weighted Avg.    0.863     0.512     0.847     0.863     0.85      0.818

=== Confusion Matrix ===

   a     b   <-- classified as
15798   770 |  a = 0
 1938  1278 |  b = 1
```

**Figura 25**: *Salida de Resultados – WEKA (Algoritmo C4.5)*
**Fuente**: *Propia de los autores*

Además[34], si se observa la diagonal de la matriz de confusión, se tiene unos valores superiores a los elementos $a_{21}$ y $a_{12}$. Esto es, $a_{11}$=15.798 es mayor que $a_{21}$=1938; y por otro lado, $a_{22}$=1.278 es mayor que $a_{12}$=770. En concreto, se observa que un 39,7% de las pólizas anuladas son clasificadas correctamente y un 95,4% de las pólizas en vigor.

Cabe mencionar el resultado del Índice Kappa, el cual mide si la concordancia establecida se debe exclusivamente al azar. Es decir, un índice *0* es el que se espera cuando la concordancia de los datos se debe al azar.

Por el contrario, si este índice es mayor que cero, se interpreta cuando los datos no son exclusivamente aleatorias. En este caso, se tienes un índice de 0.4111, indicando que la clasificación de las pólizas, como anulados o vigor, no es aleatoria.

### 4.3.2. Análisis de las Principales Ramas

Para analizar el modelo obtenido por la metodología C4.5 e interpretar dicho árbol, habría que ir descendiendo, hasta completar la totalidad de sus hojas (regla de decisión). Al final de cada hoja del árbol aparece un valor $n$ o $n/m$, siendo su interpretación:

- $n$: representa el número de pólizas en la muestra que se clasifican de acuerdo a las condiciones que nos llevan hasta esa hoja
    - $m$: representa el número de pólizas mal clasificadas

[34] *Validación-Cruzada* es un procedimiento que consiste en hacer numerosas particiones de igual tamaño en los datos, dejando unas para estimar el modelo y las restantes para validar. El proceso se repite tantas veces como particiones hayamos hecho, y vamos cambiando las que sirven para estimar de las que sirven para validar. El resultado final es la media de todos los resultados obtenidos. El resultado obtenido es para una validación cruzada para 10 particiones, que es la más habitual

Así se tiene en el primer nivel de *"ramas"* a la variable ANTIGÜEDAD (*Figura 26*):

```
=== Classifier model (full training set) ===
ANTIGUEDAD = 0: 0 (0.0)
ANTIGUEDAD = 1
|   TIPO_PRODUCTO = 1
|   |   ICE = 0: 1 (0.0)
|   |   ICE = 1: 1 (53.0/6.0)
|   |   ICE = 2
|   |   |   EDAD_ACTUARIAL <= 35: 1 (12.0)
|   |   |   EDAD_ACTUARIAL > 35
|   |   |   |   VALOR_CLIENTE = A: 0 (10.0)
|   |   |   |   VALOR_CLIENTE = B: 0 (0.0)
|   |   |   |   VALOR_CLIENTE = C: 1 (5.0/1.0)
|   |   |   |   VALOR_CLIENTE = D: 1 (14.0/5.0)
|   |   |   |   VALOR_CLIENTE = E: 0 (3.0)
|   |   |   |   VALOR_CLIENTE = F
|   |   |   |   |   SEXO = V
.....
.....(continúa)
.....
|   TIPO_PRODUCTO = 2
|   |   SEXO = V: 1 (223.0/85.0)
|   |   SEXO = H: 0 (147.0/59.0)
ANTIGUEDAD = 2
|   EDO_CIVIL = 0: 1 (0.0)
|   EDO_CIVIL = 1
|   |   NIV_ESTUDIOS = 00: 1 (0.0)
|   |   NIV_ESTUDIOS = 01
.....
.....(continúa)
.....
|   |   |   EDAD_ACTUARIAL > 48: 0 (9.0)
|   |   EDO_CIVIL = 6
|   |   |   HIJOS = N: 0 (5.0/1.0)
|   |   |   HIJOS = S: 1 (8.0/3.0)
|   |   EDO_CIVIL = 7: 0 (0.0)
|   |   EDO_CIVIL = 8: 0 (0.0)
|   |   EDO_CIVIL = 9: 0 (3.0/1.0)
|   RED = 2: 0 (62.0/11.0)
ANTIGUEDAD = 4: 0 (3851.0/430.0)
ANTIGUEDAD = 5: 0 (3362.0/208.0)
ANTIGUEDAD = 6: 0 (2760.0/81.0)
ANTIGUEDAD = 7: 0 (1836.0/93.0)
ANTIGUEDAD = 8: 0 (1570.0/47.0)
ANTIGUEDAD = 9: 0 (643.0/2.0)
ANTIGUEDAD = 10: 0 (447.0)
ANTIGUEDAD = 11: 0 (235.0)
ANTIGUEDAD = 12: 0 (48.0)
```

**Figura 26***: Salida de Resultados del Árbol de Decisión – WEKA (Algoritmo C4.5)*
**Fuente**: *Propia de los autores*

Como se puede observar esta primera variable ANTIGÜEDAD es una pieza clave en el modelo. Esto suena lógico desde el punto de vista de la existencia de cierto nivel de fidelización por parte de los clientes. Esto es, a mayor antigüedad dentro de la compañía de seguros, menor es la propensión o susceptibilidad que se tiene de cancelar su contrato de seguros.

Ahora bien, también se logra observar que existe cierto punto de inflexión a partir del cuarto -quinto año con la compañía; ya que a partir de que la variable ANTIGÜEDAD toma el valor de 4, el algoritmo de árboles no recoge alguna otra variable que nos indique cierto patrón de comportamiento de los clientes. Por el contrario, a partir de dicha cantidad de años, únicamente clasifica las pólizas no anuladas o retenidas por la entidad.

De esta forma, el problema para la compañía es que su cartera sobrepase los 4 años de antigüedad, es decir, el análisis de caída de cartera tendría que centrarse en los primeros años de vida de las pólizas, que es donde realmente se presenta el riesgo de caída de cartera y donde la compañía de seguros debería enfocar sus esfuerzos en la retención de estos clientes.

Siendo así, es necesario realizar el análisis del árbol, subdividiendo la muestra de acuerdo a ésta primera variable; que corresponde al primer nivel de *"ramas"*. Se tiene una cartera total de 19.784 pólizas, de las cuales 14.752 pólizas son clasificadas por el árbol únicamente como pólizas en VIGOR mediante la variable ANTIGÜEDAD, lo cual nos refiere que dicha variable es un factor clave en el comportamiento de los clientes.

Partiendo de esto, se tiene que el 75% de la muestra, cuenta con más de 4 años de antigüedad dentro de la compañía, lo cual se justifica principalmente por la fidelización de los clientes hacia la marca; sin poder distinguir otro tipo de variable o característica del asegurado que determine específicamente el tipo de clientes que se retienen en cartera.

Es así como se debe substraer un tipo "*subcartera*" correspondiente a la cartera de clientes más recientes (*Figura 27*):

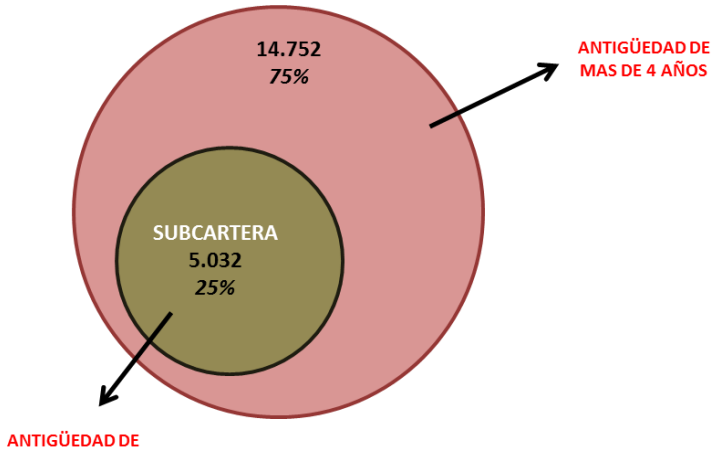


**Figura 27**: *Composición de la cartera para la subdivisión de los análisis de acuerdo a la variable ANTIGUEDAD*
**Fuente**: *Propia de los autores*

Así pues poder tener una fuerza global muy alta para la cartera total explicada principalmente por la variable ANTIGÜEDAD; y complementarla con una fuerza relativa determinada con respecto a la *subcartera*. De esta forma, se centrará el análisis en los patrones de las antigüedades 1, 2 y 3 que corresponde al 25% de la muestra (5.032 pólizas), cuyo comportamiento de los asegurados dentro de esta *subcartera* se podría considerar independiente al resto de la muestra.

### 4.3.3. Análisis de los Principales Patrones de las Pólizas Recientes

Es así como se centra el análisis de los patrones sobre la *subcartera* de pólizas más recientes de la compañía cuya ANTIGÜEDAD se encuentra entre 1 y 3 años de duración. A partir de esta *subcartera*, se obtienen las principales reglas de decisión sobre los 2 posibles comportamientos del asegurado:

- *TIPO PRESTACION = 1* correspondiente a los factores determinantes de los clientes susceptibles a la cancelación de su póliza; y

- *TIPO PRESTACION = 0* correspondiente a los patrones presentados en los clientes propensos a mantener su contrato de seguros en vigor

#### 4.3.3.1. Arboles de Decisión correspondientes a la CLASE 1=Cancelación

Para comprender mejor el árbol de decisión obtenido, se puede interpretar gráficamente. Para ello, se han identificado algunas de las *ramas* correspondientes a las reglas con mayor fuerza. A continuación se presentan algunos de estas *ramas* del árbol que dan lugar a las principales reglas detectadas para la CLASE 1; esto es, las reglas de decisión que determinarían el tipo de clientes propensos a la cancelación de su póliza de seguros:

- REGLA 1 DE LA CLASE 1=CANCELACION

Siendo así, una primera Regla detectada se muestra gráficamente (*Figura 28*); la cual se describe cada una de las *ramas* del árbol de decisión de la siguiente forma:

- La primera rama que se tiene en cuenta es la ANTIGÜEDAD. Todas las pólizas, según el criterio de TIPO DE PRESTACION, se pueden clasificar atendiendo, en primer lugar, a la ANTIGÜEDAD, es decir, los años de antigüedad de la póliza desde su Fecha de Emisión hasta la Fecha de Cálculo. Dicha duración puede ir desde 1 año hasta 12 años de antigüedad con la entidad aseguradora. En particular, para esta porción del árbol seleccionado, se analiza los casos que cumplen con tener 1 año de duración dentro de la compañía; las cuales requieren del análisis de otra variable adicional, el TIPO PRODUCTO.

- Así se tiene que, la segunda rama es el TIPO PRODUCTO contratado, esto es, se trata de una póliza de *Vida Riesgo,* si la variable es igual a 2; o *Vida Ahorro,* si toma el valor de 1. Pues bien, considerando este nivel, donde el TIPO PRODUCTO

contratado es de *Vida Ahorro,* se continúa bajando a la tercera rama del árbol de decisión.

- De esta forma, se llega a la tercera rama del árbol que es la variable ICE. Esta variable puede tomar distintos valores: *1, 2,3,* etc.; que corresponde a los distintos niveles como se ha categorizado el Índice de Capacidad Económica. Ahora bien, debido a la fuerza baja que toman algunos de sus valores, por ejemplo cuando el ICE=0 ó ICE=1; no se toman en cuanta como reglas fuertes de patrones. Lo mismo sucede cuando el ICE=2, 4 ó 5; ya que estos casos, el árbol continúa ramificándose pero bajando por el resto de ramas, no se logra encontrar ninguna regla con una fuerza significativa. La única rama donde sí se puede encontrar cierta fuerza significativa es en el caso del ICE=3, el cual se trata a un nivel MEDIO. Siendo así, se procede al siguiente nivel del árbol.

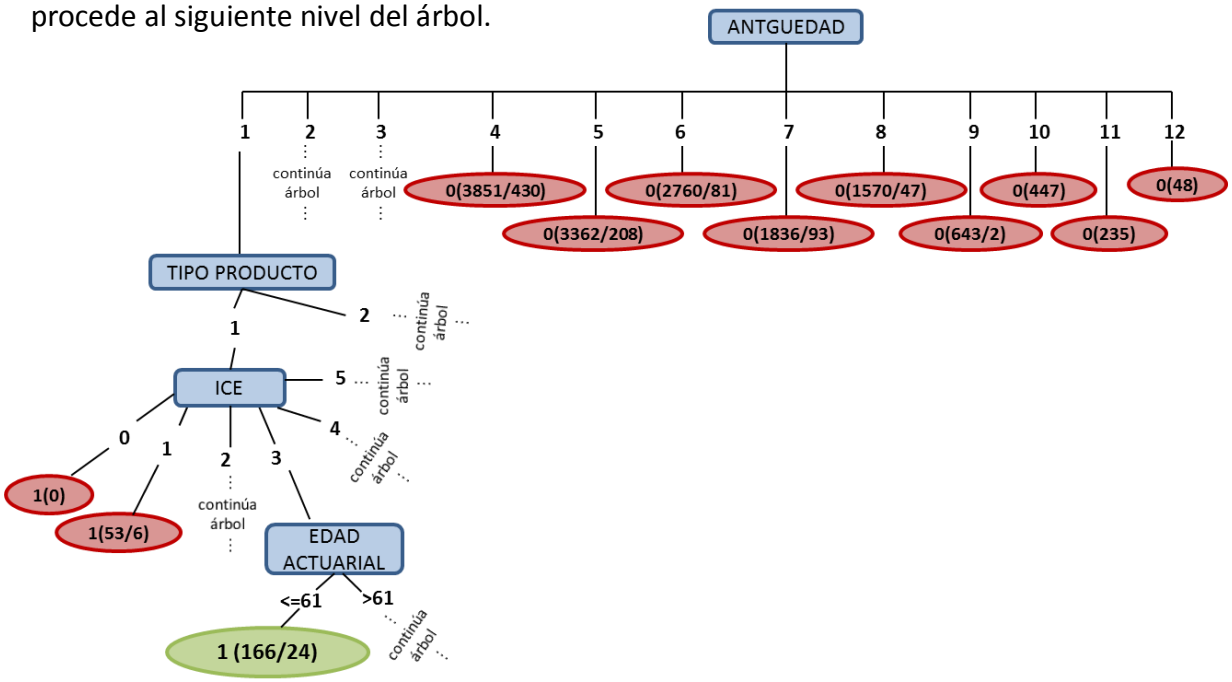


**Figura 28:** *Árbol. De Decisión (Ramo de ANTIGÜEDAD = 1 _ Regla 1 de la CLASE 1*
**Fuente**: *Propia de los autores*

- Esta última rama se obtiene la variable EDAD ACTUARIAL, la cual corresponde a los rangos de Edad en las que se ha agrupado la muestra. Así, se destaca el grupo de *<=61 años*, como el perfil del cliente susceptible a anular su póliza de seguros. Finalmente, en términos matemáticos, en conjunto con las 4 ramas anteriores, este patrón **se cumple en 166 de los casos, esto es en un total de 3.3%**

**como una fuerza relativa, ya que se está considerando sobre el total de pólizas de la *subcartera*.**

➢ REGLA 2 DE LA CLASE 1=CANCELACION

Ahora bien, se tiene una segunda Regla detectada que se describe a continuación y se muestra gráficamente (*Figura 29*):

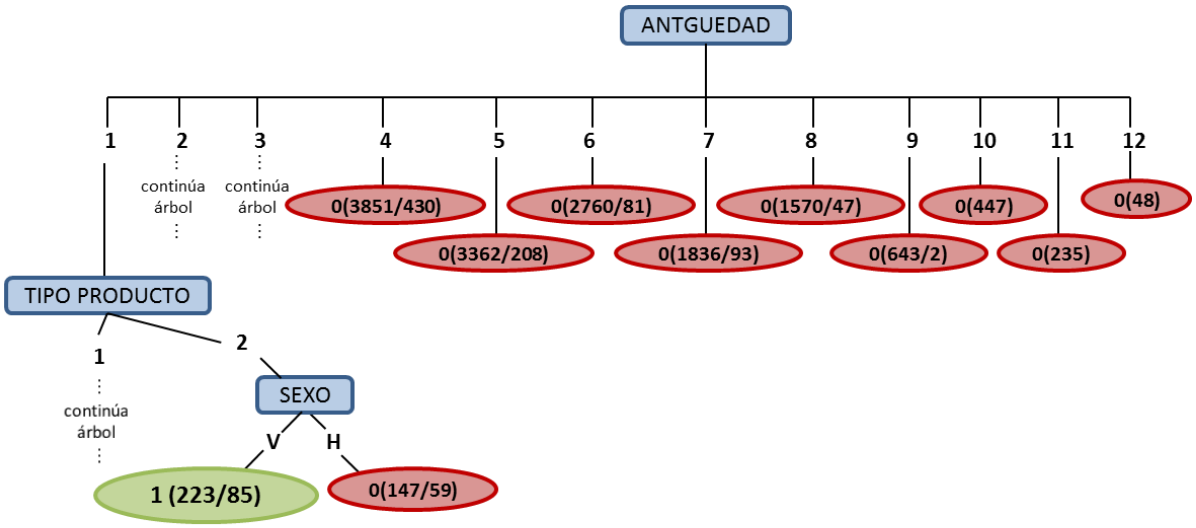


**Figura 29**: *Árbol. De Decisión (Ramo de ANTIGÜEDAD = 1)_ Regla 2 de la CLASE 1*
**Fuente**: *Propia de los autores*

- La primera rama que se tiene en cuenta nuevamente es la ANTIGÜEDAD, es decir, los años de antigüedad de la póliza desde su Fecha de Emisión hasta la Fecha de Cálculo. Una vez más, se analiza los casos que cumplen con tener 1 año de duración dentro de la compañía. Siendo así, se requiere descender nuevamente hacia la siguiente rama del árbol

- Así se pasa a la segunda rama del árbol que corresponde a la variable TIPO PRODUCTO contratado. En este caso, se continúa sobre la ramificación cuando la variable toma el valor de 2; es decir, se trata de una póliza de *Vida Riesgo*. Pues bien, considerando de esta forma dicho nivel, se continúa bajando a la tercera rama del árbol de decisión

- De esta forma, se llega a la tercera rama que corresponde a la variable SEXO; la cual toma el valor de *H* que engloba el grupo de las *Mujeres*; o bien *V* cuando se trata del grupo de los *Hombres*. Pues bien, cuando la variable toma este último

valor, se llega a una segunda de regla de decisión con una fuerza significativa para identificar al tipo de clientes susceptibles a anular su póliza de seguros. En términos matemáticos, en conjunto con las 2 ramas anteriores, este patrón **se cumple en 223 de los casos, esto es en un total de 4.4% como una fuerza relativa sobre el total de pólizas de la *subcartera*.**

➢ REGLA 3 DE LA CLASE 1=CANCELACION

De la misma forma, se tiene una tercera Regla de Decisión que se describe a continuación y se muestra gráficamente (*Figura 30*):

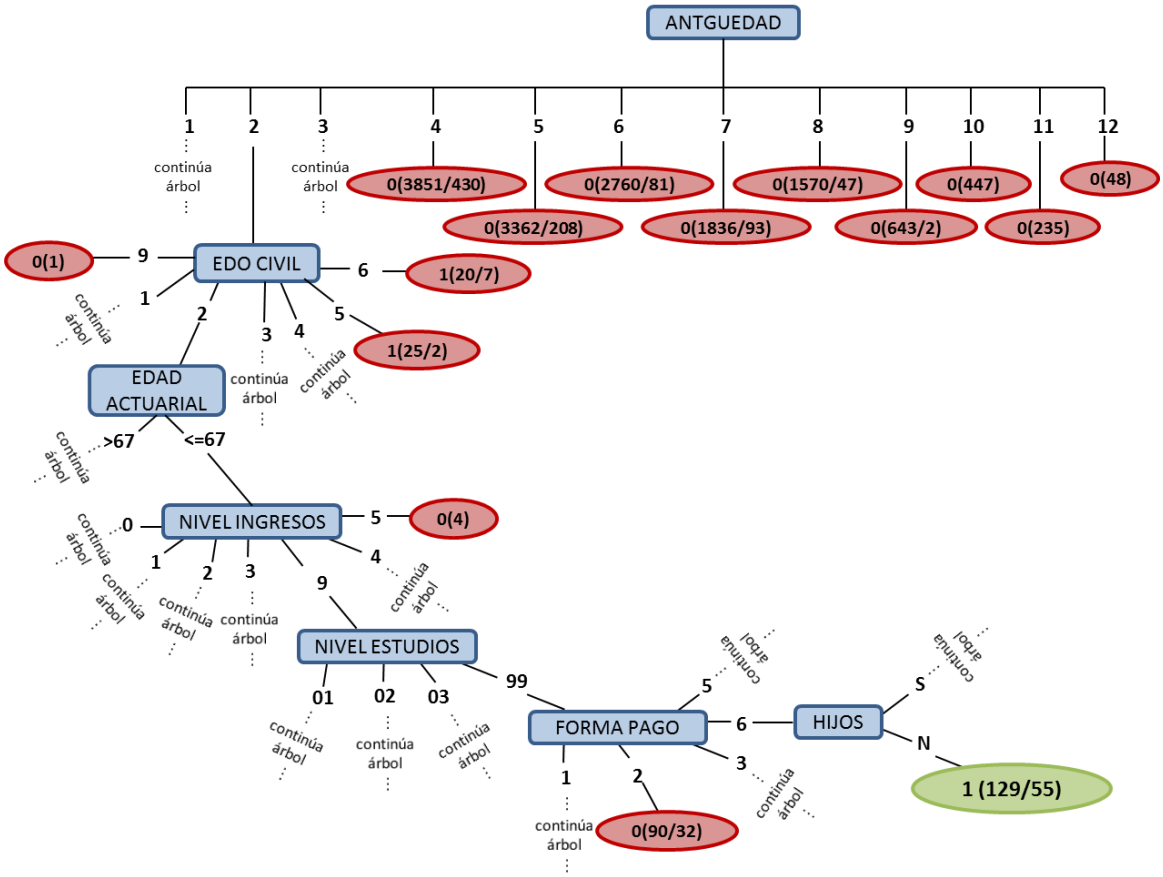


**Figura 30:** *Árbol. De Decisión (Ramo de ANTIGÜEDAD = 2_ Regla 3 de la CLASE 1*
**Fuente**: *Propia de los autores*

- La primera rama se comparte nuevamente siendo la variable ANTIGÜEDAD, pero ahora tomando el valor de 2; es decir, esta vez considerando las pólizas que ya han durado 2 años dentro la entidad. Se procede al siguiente nivel descendiendo por el árbol

- Así se llega a la segunda rama del árbol que corresponde esta ocasión a la variable ESTADO CIVIL. Esta variable puede tomar distintos valores: *1, 2, 3, 4,* etc.; que corresponde a los distintos Estados Civiles declarados: *Soltero, Casado, Divorciado, Viudo,* etc. Ahora bien, debido a que el volumen de cartera que queda segmentada por cada una de dichos valores es poco significativa, es preciso considerar únicamente la variable ESTADO CIVIL cuando toma el valor 2, esto es *Casado*

- Continuando por el árbol, se desciende a la tercera rama que corresponde a la variable EDAD ACTUARIAL, donde nuevamente corresponde a los rangos de Edad en las que se ha agrupado la muestra. En esta ocasión, se destaca el grupo de *<=67 años*, como la rama por donde se continúa el análisis

- Se llega a la cuarta variable NIVEL INGRESOS con la cual el árbol se ramifica debido a los distintos valores que puede tomar esta variable: *1, 2, 3, 4,* etc.; que corresponde a los distintos rangos en que se han agrupado los Niveles de Ingresos declarados: *<6.000€, De 6.000€ a 18.000€, De 18.000€ a 36.000€,* etc. En este caso, se considera únicamente la variable NIVEL INGRESOS cuando toma el valor 9, esto es *No Declarado/Sin Información*

- Bajando por esta rama se tiene una quinta variable NIVEL ESTUDIOS, la cual toma 4 valores: *01, 02, 03, 99*, que corresponde a la categorización del Nivel de Estudios declarado por el cliente, siendo: *Elementales, BUP/PP/ESO, UNIVERSITARIOS y Sin Informar*. En este caso, se considera la variable NIVEL ESTUDIOS cuando toma el valor *99*; lo que corresponde al grupo de clientes que no han informado su nivel estudios. Este resultado, sugiere que existen pocas pólizas que realmente declaran tanto, el Nivel de Ingresos, como el Nivel de Estudios, que tienen. Por lo que se sugiere "obviar" ambas ramas; es decir, obviar ambas variables dentro de la regla de decisión obtenida. Siendo así, se procede al siguiente nivel del árbol.

- Así se tiene que una sexta rama del árbol que es la FORMA PAGO; la cual toma los valores: *1, 2, 3,* etc.; correspondiente a las distintas Formas de Pago con las que cuenta el asegurado para hacer el pago de su contrato: *Anual, Semestral, Mensual*, etc. En este caso, se procede a descender por la rama con mayor fuerza; esto es,

cuando la variable toma el valor de 6, es decir, una Forma de Pago Única de la póliza de seguros

- Sobre esta ramificación, se llega a la última rama teniendo la variable HIJOS; la cual puede tomar el valor de *N* que significa que *No Hijos*; o bien *S* cuando se declara *Si Hijos*. Finalmente, cuando la variable toma el valor de *No Hijos*, se llega a una tercera regla de decisión para identificar al tipo de clientes susceptibles a anular su póliza de seguros. En términos matemáticos, en conjunto con todas las ramas anteriores, este patrón **se cumple en 129 de los casos, esto es en un total de 2.6% como una fuerza relativa sobre el total de pólizas de la *subcartera*.**

#### 4.3.3.2. Arboles de Decisión correspondientes a la CLASE 0=Retención

Por otro lado se tiene un segundo conjunto de reglas detectadas para la CLASE 0; es decir patrones de comportamiento que determinan el tipo de clientes susceptibles a retener o conservar su póliza de seguros. Para ello, de igual forma, se puede interpretar gráficamente algunas de las *ramas* correspondientes a las reglas con mayor fuerza para la CLASE 0:

- <u>REGLA 1 DE LA CLASE 0=RETENCION</u>

Siendo así, una primera Regla detectada se muestra gráficamente (*Figura 31*); la cual se describe cada una de las *ramas* del árbol de decisión de la siguiente forma:

- La primera rama que se tiene en cuenta es la ANTIGÜEDAD. Nuevamente, se desciende por el árbol cuando dicha duración toma el valor de 1 año de antigüedad con la entidad aseguradora. Siendo así, se requiere descender hacia la siguiente rama del árbol

- Así se pasa a la segunda rama del árbol que corresponde a la variable TIPO PRODUCTO contratado. Una vez más se continúa sobre la ramificación cuando la variable toma el valor de 2; es decir, se trata de una póliza de *Vida Riesgo*

- Continuando descendiendo, se llega a la tercera rama que corresponde a la variable SEXO; la cual engloba al grupo de las *Mujeres*; o bien al grupo de los *Hombres*. Pues bien, en este caso, se considera cuando la variable toma el valor de *H* correspondiente al grupo femenino. De esta forma, se llega a una primera regla de decisión sobre la Clase de Retención, es decir el patrón que identifica al tipo de clientes susceptibles a conservar su póliza de seguros. En términos matemáticos, en conjunto con las 2 ramas anteriores, este patrón **se cumple en 147 de los casos, esto es en un total de 2.9% como una fuerza relativa sobre el total de pólizas de la *subcartera*.**

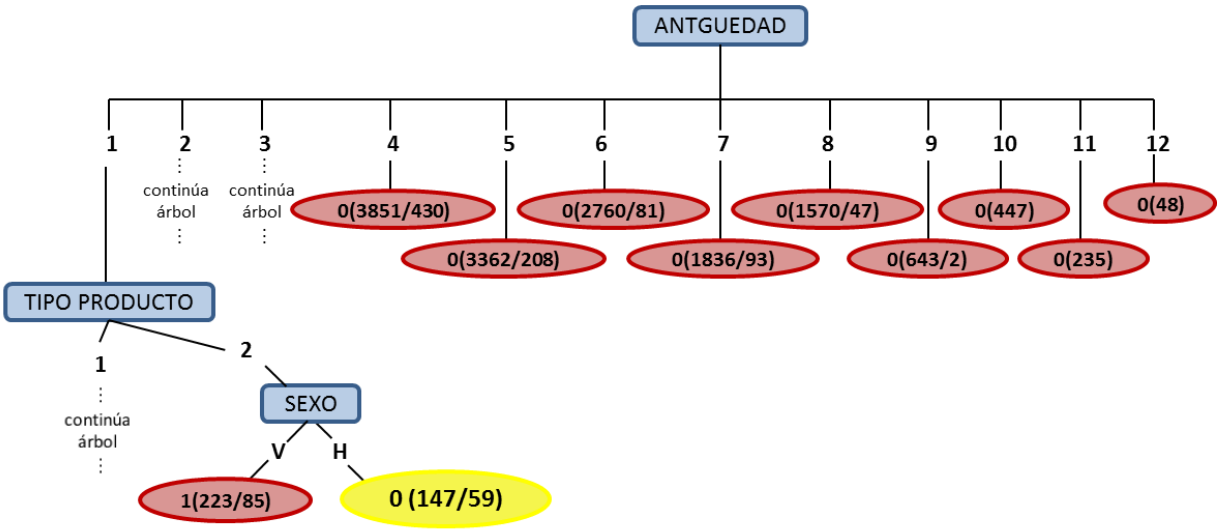


**Figura 31:** *Árbol. De Decisión (Ramo de ANTIGÜEDAD = 1)_ Regla 1 de la CLASE 0*
**Fuente**: *Propia de los autores*

➢ REGLA 2 DE LA CLASE 0=RETENCION

De la misma forma, se tiene una segunda Regla de Decisión que se describe a continuación y se muestra gráficamente (*Figura 32*):

- La primera rama se comparte nuevamente la variable ANTIGÜEDAD, pero ahora tomando el valor de 2; es decir, considerando las pólizas que ya han durado 2 años dentro la entidad

- Se baja a la segunda rama del árbol que corresponde a la variable ESTADO CIVIL. Esta variable puede tomar distintos valores: *1, 2, 3, 4,* etc.; que corresponde a los distintos Estados Civiles declarados: *Soltero, Casado, Divorciado, Viudo,* etc.; sin embargo debido al poco volumen de la cartera segmentada por los distintos valores, es

preciso considerar únicamente la variable ESTADO CIVIL cuando toma el valor 2, esto es *Casado*

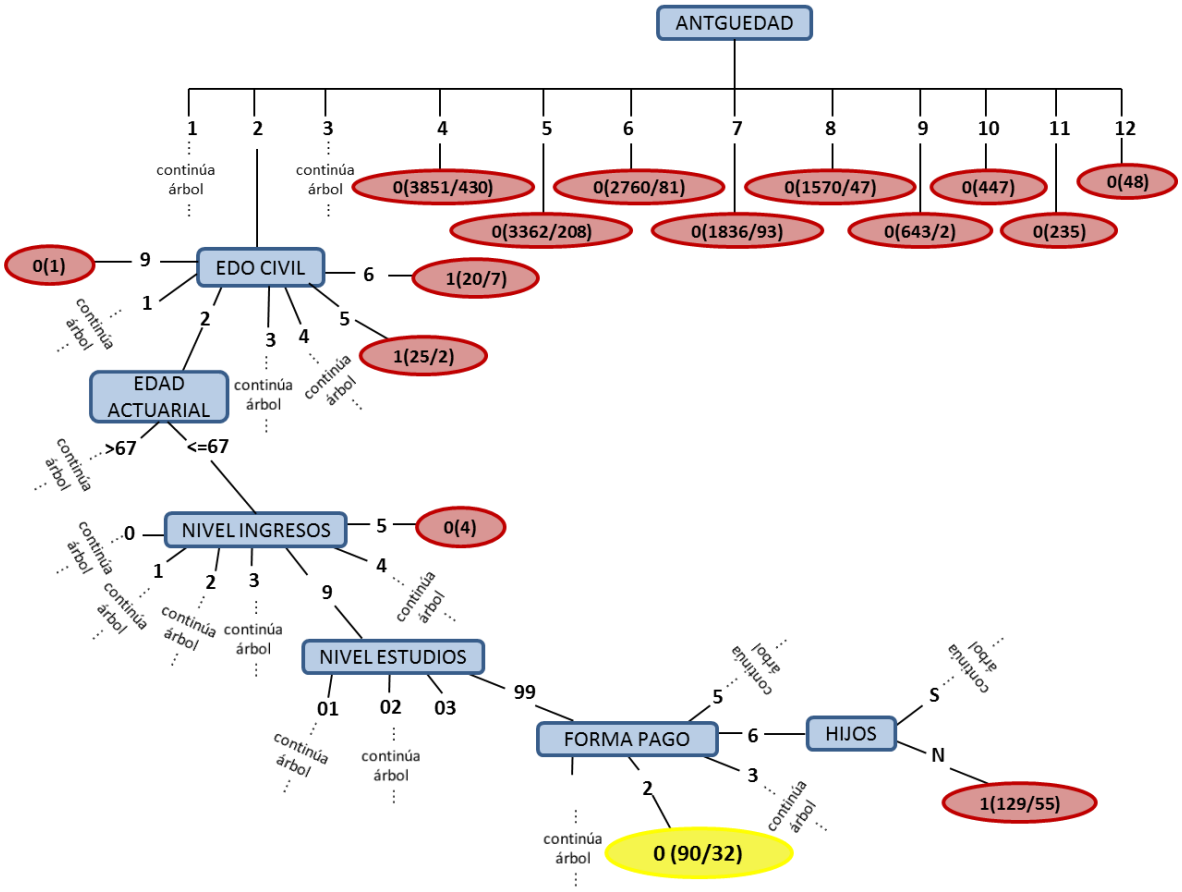


**Figura 32**: *Árbol. De Decisión (Ramo de ANTIGÜEDAD = 2_ Regla 2 de la CLASE 0*
**Fuente**: *Propia de los autores*

- Se desciende a la tercera rama que corresponde a la variable EDAD ACTUARIAL, donde nuevamente corresponde a los rangos de Edad en las que se ha agrupado la muestra. En esta ocasión, se destaca el grupo de *<=67 años*, como la rama por donde se continúa descendiendo

- Se llega a la cuarta variable NIVEL INGRESOS con la cual el árbol se ramifica debido a los distintos valores que corresponde a los distintos rangos en que se han agrupado los Niveles de Ingresos declarados. Una vez más, se considera únicamente la variable NIVEL INGRESOS cuando toma el valor 9, esto es *No Declarado/Sin Información*

- Continuando por esta rama se tiene una quinta variable NIVEL ESTUDIOS, la cual toma 4 valores que corresponde a la categorización del Nivel de Estudios

declarado por el cliente. Nuevamente, se desciende por la rama donde la variable NIVEL ESTUDIOS corresponde al grupo de clientes que no han informado su nivel estudios. Únicamente, comentar que se sugiere "obviar" tanto esta rama como la anterior de la regla de decisión obtenida; ya que existen pocas pólizas que realmente declaran tanto, el Nivel de Ingresos, como el Nivel de Estudios, que tienen

- Se llega a la última rama del árbol que es la FORMA PAGO; la cual toma los valores correspondientes a las distintas Formas de Pago con las que cuenta el asegurado para hacer el pago de su contrato. Es así como se llega a la segunda regla de decisión para identificar al tipo de clientes susceptibles a retener su póliza de seguros, cuando dicha variable toma el valor de *2,* es decir, cuando se trata de una Forma de Pago *Semestral*. Una vez más, este patrón en términos matemáticos, en conjunto con todas las ramas anteriores, **se cumple en 90 de los casos, esto es en un total de 1.8% como una fuerza relativa sobre el total de pólizas de la *subcartera*.**

- REGLA 3 DE LA CLASE 0=RETENCION

Continuando, se puede obtener una tercer Regla de Decisión que se describe en seguida y se muestra gráficamente (*Figura 33*):

- La primera rama se comparte nuevamente la variable ANTIGÜEDAD, pero ahora tomando el valor de 3; es decir, considerando las pólizas que ya han durado 3 años dentro la compañía de seguros

- Se desciende a una segunda rama del árbol que corresponde a la variable RED. Esta variable puede tomar dos valores: *1* o *2*; que corresponde a los dos tipos de Redes de Distribución de las pólizas de seguros vendidas por la entidad: *Propietaria* y *No Propietaria*. En este caso, se considera esta variable cuando toma el valor 1, es decir *Red Propietaria*

- Se procede a bajar a una tercera rama del árbol hacia la variable ESTADO CIVIL. Nuevamente esta variable puede tomar los valores correspondientes a los distintos Estados Civiles declarados: *Soltero, Casado, Divorciado, Viudo,* etc. Así se llega a la tercera regla de decisión para identificar al tipo de clientes susceptibles a retener

su póliza de seguros, cuando dicha variable toma el valor de *1,* es decir, cuando se trata de un *Soltero*. Este patrón en términos matemáticos, en conjunto con las ramas anteriores, **se cumple en 211 de los casos, esto es en un total de 4.2% como una fuerza relativa sobre el total de pólizas de la *subcartera*.**

➢ REGLA 4 DE LA CLASE 0=RETENCION

Se procede a obtener una cuarta Regla de Decisión que se describe a continuación y se muestra gráficamente (*Figura 34*):

- La primera rama se inicia nuevamente con la variable ANTIGÜEDAD, pero ahora tomando las pólizas que ya han durado 3 años dentro la compañía de seguros

- Se baja, una vez más, hacia la variable RED como una segunda rama del árbol; considerando esta variable cuando toma el valor de *Red Propietaria*

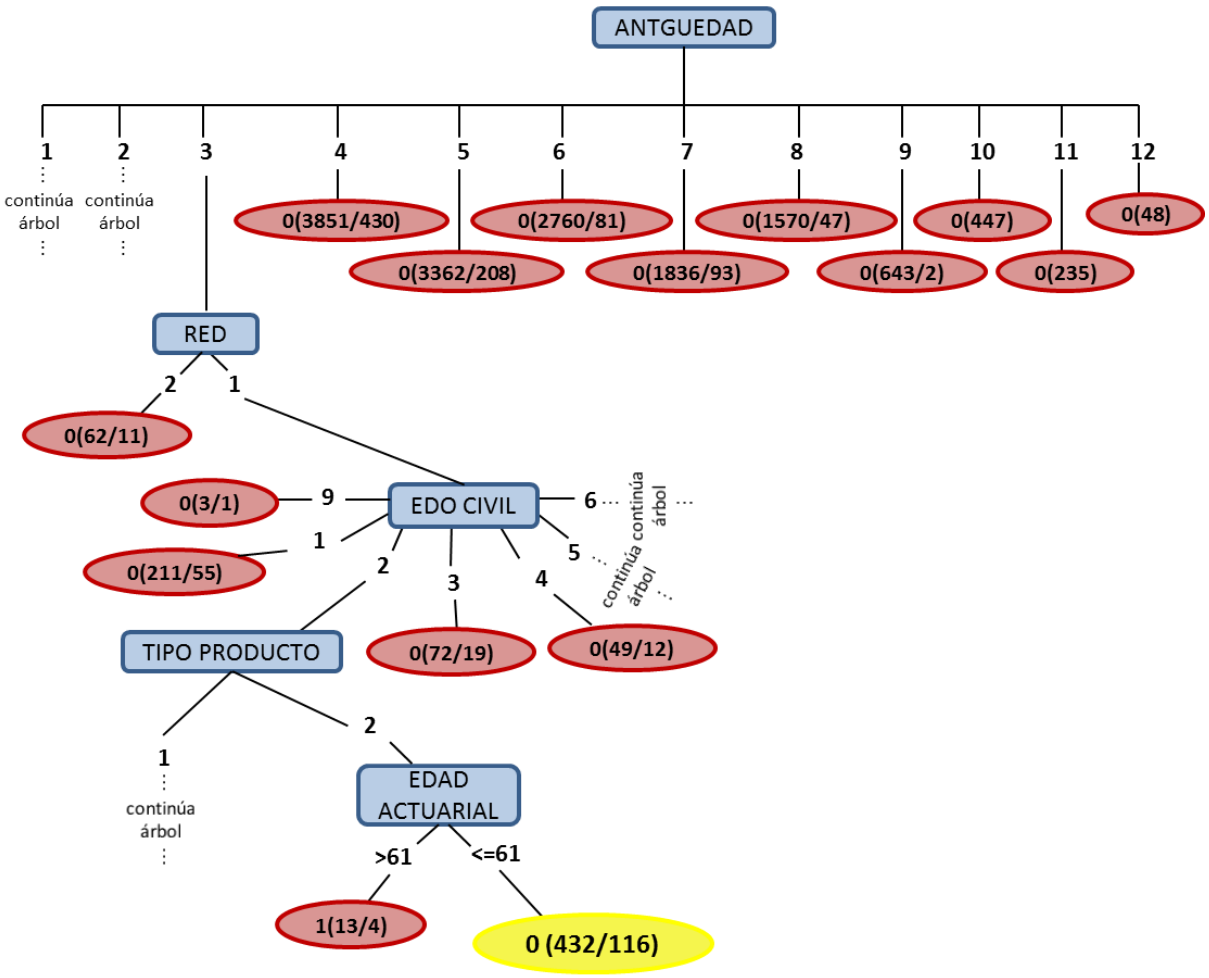


**Figura 34***: Árbol. De Decisión (Ramo de ANTIGÜEDAD = 3 _ Regla 4 de la CLASE 0*
**Fuente**: *Propia de los autores*

- Se desciende hacia la variable ESTADO CIVIL, que como ya se comentó, esta variable puede tomar distintos valores: *Soltero, Casado, Divorciado, Viudo,* etc. En esta ocasión, se sigue el árbol cuando toma el valor de 2 correspondiente al Estado Civil *Casado;* presentando la necesidad de continuar hacia una siguiente rama

- Así se llega a una cuarta rama correspondiente a la variable TIPO PRODUCTO contratado; continuando sobre la ramificación cuando la variable toma el valor de 2; es decir, se trata de una póliza de *Vida Riesgo*

- Se llega a la última rama que corresponde a la variable EDAD ACTUARIAL, obteniendo así la cuarta regla de decisión para identificar al tipo de clientes susceptibles a retener su póliza de seguros, cuando la variable EDAD ACTUARIAL es menor o igual a 61 años. Este patrón en términos matemáticos, en conjunto con las ramas anteriores, **se cumple en 432 de los casos, esto es en un total de 8.6% como una fuerza relativa sobre el total de pólizas de la *subcartera*.**

### 4.3.4. Principales Reglas de Decisión de las Pólizas Recientes

Una vez interpretados las ramificaciones más destacadas del árbol, se puede resumir algunas de Reglas de Decisión. Esto es, en otras palabras, el conjunto de patrones o variables que ayudarían a clasificar a los clientes hacia una determinado tipo de comportamiento.

Así pues, primeramente, como ya se ha mencionado, existe una fuerza significativa explicada principalmente por la variable ANTIGÜEDAD, siendo una primera variable que clasifica a los clientes susceptibles a conservar su póliza de seguros. Esto es, los clientes que cuentan con más de 4 años de antigüedad con la entidad, son altamente propensos a no cancelar su póliza, presentándose esta regla en un 75% de la muestra (*cartera total* = 19.784 pólizas).

Ahora bien, para el otro 25% de la muestra, a la cual se ha denominado *subcartera*, se han encontrado una serie de reglas para el tipo de pólizas "*recientes*",

que se presentan a continuación para cada una de las clases: Cancelación o Retención (*subcartera* = 5.032 pólizas).

#### 4.3.4.1. Reglas de Decisión correspondientes a la CLASE 1=Cancelación

A continuación se presentan las principales reglas detectadas para la CLASE 1; esto es, las reglas de decisión que determinarían el tipo de clientes propensos a la cancelación de su póliza de seguros, sumando un total de **10.3% de fuerza relativa medida sobre el número de póliza que conforman la *subcartera de pólizas "recientes"***:

- REGLA 1 DE LA CLASE 1=CANCELACION

La primera Regla detectada sería:

***ANTIGÜEDAD=1 → TIPO PRODUCTO=1 → ICE=3 → EDAD ACTUARIAL<=61***

Es decir que, aquellos clientes con antigüedad de 1 año, cuya póliza contratada es de Vida Ahorro, con un Índice de Capacidad Económica Medio y con Edad por debajo de los 61 años; tienden a cancelar su contrato de seguros y abandonar la compañía de seguros.

Esto se cumple en:

- En **166** casos
- Representado un **3.3%** sobre el número de pólizas de la subcartera
- Y en un **0.8%** sobre la cartera total

- REGLA 2 DE LA CLASE 1=CANCELACION

La segunda Regla detectada sería:

***ANTIGÜEDAD=1 → TIPO PRODUCTO=2 → SEXO=V***

En este caso se trata de aquellos clientes con antigüedad de 1 año, cuya póliza contratada es de Vida Riesgo y de Sexo Masculino quienes marcan un segundo patrón de comportamiento de los clientes susceptibles a anular su póliza de seguros.

Esto cumpliéndose en:

- En **223** casos
- Representado un **4.4%** sobre el número de pólizas de la subcartera
- Y en un **1.1%** sobre la cartera total

- REGLA 3 DE LA CLASE 1=CANCELACION

Una tercera Regla detectada para esta clase sería:

***ANTIGÜEDAD=2 → ESTADO CIVIL=2 → EDAD ACTUARIAL<=67 → NIVEL INGRESOS=9 → NIVEL ESTUDIOS=99 → FORMA PAGO=6 → HIJOS=N***

En este caso se trata de aquellos clientes con 2 años de antigüedad, cuyo Estado Civil declarado es Casado, con una Edad menor a 67 años, cuyo Nivel de Ingresos y Nivel de Estudios no ha sido declarado (por lo que se podría obviar dicho característica), con una Forma de Pago de su póliza es mediante un Pago único y finalmente, que ha declarado No tener Hijos. Este grupo de clientes muestran la tercera regla de comportamiento de los clientes susceptibles a cancelar su contrato de seguros.

Esto observándose en:

- En **129** casos
- Representado un **2.6%** sobre el número de pólizas de la subcartera
- Y en un **0.7%** sobre la cartera total

De esta forma, el resumen de reglas para la Clase 1=Cancelación, sumando estas 3 reglas acumulan una fuerza del 2.6% sobre la cartera total de pólizas; y a su vez, representa un 10.3% como fuerza relativa sobre la *subcartera (Tabla 28)*:

**RESUMEN REGLAS**

| Categoría: 1 |
|---|
| CAIDA |

| CARTERA | SUBCART |
|---|---|
| 19,784 | 5,032 |

| | | | | | | | | s / CARTERA | s / SUBCART |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 518 | *2.6%* | *10.3%* |
| REGLA 1 | PROD - Ahorro | ICE - Medio | EDAD - <=61 | | | | 166 | *0.8%* | *3.3%* |
| REGLA 2 | PROD - Riesgo | SEXO - Hombre | | | | | 223 | *1.1%* | *4.4%* |
| REGLA 3 | EDO - Casado | EDAD - <=67 | INGRS - na | ESTUD - na | FP - Unica | Hijos N | 129 | *0.7%* | *2.6%* |

**Tabla 28**: *Resumen de Resultados Arboles de Decisión – CLASE 1: CAIDA*
**Fuente**: *Propia de los autores*

### 4.3.4.2. Reglas de Decisión correspondientes a la CLASE 0=Retención

Ahora bien, las principales reglas detectadas para la CLASE 0; que determinarían el tipo de clientes propensos a la conservación de su póliza de seguros, sumando un total de **17.5% de fuerza relativa medida sobre el número de póliza que conforman la *subcartera de pólizas "recientes"***:

- REGLA 1 DE LA CLASE 0=RETENCION

La primera Regla detectada para la Clase de Retención sería:

***ANTIGÜEDAD=1 → TIPO PRODUCTO=2 → SEXO=H***

En este caso, se trata de aquellos clientes con antigüedad de 1 año, cuya póliza contratada es de Vida Riesgo y de Sexo Femenino quienes marcan un primer patrón de comportamiento de las personas propensas a conservar su póliza de seguros.

Esto se cumple en:

- En **147** casos
- Representado un **2.9%** sobre el número de pólizas de la subcartera
- Y en un **0.7%** sobre la cartera total

➢ REGLA 2 DE LA CLASE 0=RETENCION

La segunda Regla detectada sería:

***ANTIGÜEDAD=2 → ESTADO CIVIL=2 → EDAD ACTUARIAL<=67 → NIVEL INGRESOS=9 → NIVEL ESTUDIOS=99 → FORMA PAGO=2***

En este caso se trata de aquellos clientes con 2 años de antigüedad, cuyo Estado Civil declarado es Casado, con una Edad menor a 67 años, cuyo Nivel de Ingresos y Nivel de Estudios no ha sido declarado (por lo que se podría obviar dicho característica), y con una Forma de Pago Semestral. Este conjunto muestra la segunda regla de comportamiento de los clientes propensos a conservar su contrato de seguros.

Esto cumpliéndose en:

- En **90** casos
- Representado un **1.8%** sobre el número de pólizas de la subcartera
- Y en un **0.5%** sobre la cartera total

➢ REGLA 3 DE LA CLASE 0=RETENCION

Una tercera Regla detectada para esta clase sería:

***ANTIGÜEDAD=3 → RED=1 → ESTADO CIVIL=1***

Así se tiene a los clientes con 3 años de antigüedad, cuya venta de póliza proviene de una Red Propietaria, es decir Agentes o empleados propios de la entidad;

y cuyo Estado Civil declarado es Soltero muestran la tercera regla de comportamiento de los clientes susceptibles a mantener su contrato de seguros.

Esto observándose en:

- En **211** casos
- Representado un **4.2%** sobre el número de pólizas de la subcartera
- Y en un **1.1%** sobre la cartera total

➢ REGLA 4 DE LA CLASE 0=RETENCION

Una cuarta Regla detectada para esta clase sería:

***ANTIGÜEDAD=3 → RED=1 → ESTADO CIVIL=2 → TIPO PRODUCTO=2 → EDAD ACTUARIAL<=61***

Así se tiene a los clientes cuya antigüedad es de 3 años, cuya venta de póliza proviene de una Red Propietaria, cuyo Estado Civil declarado es Casado, cuentan con una póliza de Vida Riesgo y son menores a 61 años de Edad; engloban al cuarto patrón de comportamiento de los clientes susceptibles a mantener su contrato de seguros.

Esto observándose en:

- En **432** casos
- Representado un **8.6%** sobre el número de pólizas de la subcartera
- Y en un **2.2%** sobre la cartera total

De esta forma, el resumen general de reglas para la Clase 0=Retención, englobando 4 reglas acumulan una fuerza del 4.4% sobre la cartera total de pólizas; y a su vez, representa un 17.5% como fuerza relativa sobre la *subcartera (Tabla 29)*:

**RESUMEN REGLAS**

| Categoría: 0 |
|---|
| RETENCION |

| CARTERA | SUBCART |
|---|---|
| 19,784 | 5,032 |

| | | | | | | | s / CARTERA | s / SUBCART |
|---|---|---|---|---|---|---|---|---|
| | | | | | | 880 | *4.4%* | *17.5%* |
| REGLA 1 | PROD - Riesgo | SEXO - Mujer | | | | 147 | *0.7%* | *2.9%* |
| REGLA 2 | EDO - Casado | EDAD - <=67 | INGRS - na | ESTUD - na | FP - Sem | 90 | *0.5%* | *1.8%* |
| REGLA 3 | RED - Prop | EDO - Soltero | | | | 211 | *1.1%* | *4.2%* |
| REGLA 4 | RED - Prop | EDO - Casado | PROD - Riesgo | EDAD - <=61 | | 432 | *2.2%* | *8.6%* |

**Tabla 29**: *Resumen de Resultados Arboles de Decisión – CLASE 0: RETENCION*
**Fuente**: *Propia de los autores*

### 4.3.5. Principales Resultados Obtenidos bajo Arboles de Decisión

A raíz de la aplicación de la técnica que ofrecen los Árboles de Decisión, se pueden resumir los principales resultados obtenidos. Para ello, se debe retomar el objetivo inicial de dicha aplicación; el cual es detectar una serie de patrones o variables que definen el perfil del asegurado susceptible a la cancelación de su póliza.

Es así como de los resultados que se han obtenidos, se puede decir que algunas de las variables que definen al perfil del cliente "cancelador" son, primeramente, la ANTIGÜEDAD de la póliza. Esto responde al tema de la fidelización del cliente hacia la entidad aseguradora. Ahora bien, otra de las variables que se encuentra relacionada con la duración del contrato es el TIPO PRODUCTO; el cual resulta ser un segundo patrón de comportamiento identificado. Esto habla de que se deberían tomar distinto tipo de medidas de control de caída o anulación dependiendo del tipo de póliza contratada.

Hasta este punto, podría coincidir con el tipo de variables generalmente utilizadas para el análisis de anulaciones dentro de una entidad. Sin embargo, otras

tres variables detectadas para ambas categorías analizadas, tanto Retención como Cancelación; son la EDAD y ESTADO CIVIL, con lo cual esto puede de ser de vital importancia en la toma de decisión para la contención del riesgo de caída de cartera. A la vista de estos resultados, se podrían diseñar campañas de retención hacia grupos de clientes "preferentes" de acuerdo a su Edad o bien a su Estado Civil. Es decir, sabiendo que existe cierta propensión a la anulación de la póliza de este tipo de clientes, lograr evaluar los niveles de rentabilidad que cada grupo de clientes proporciona conociendo dicha tendencia hacia el abandono de su contrato.

En cuanto a la FORMA PAGO sugiere ser otra característica interesante que resulta del modelo; ya que esto habla de motivar a los clientes a elegir cierta forma de pago de sus primas de acuerdo a la rentabilidad que ofrecen; y por lo tanto, proponer supuestos de anulación futura de la cartera diferenciados por las diferentes forma de pago que se tengan en la entidad. Así mismo, se ha obtenido que la variable SEXO sea otra de las cuestiones a tener en cuenta para evaluar la posibilidad de cancelación o retención de la póliza en un cliente. Esto, si bien ha dejado de tener consideración por temas de diversidad de género y discriminación; no deja de perder valor a la vista de resultados como éstos.

Otra de las variables con menor fuerza pero no por ello poco significativas es la variable HIJOS, es decir, si se ha declarado tener o no Hijos; que podría verse relacionada con el Estado Civil del cliente, ya que el efecto puede ser similar tomando en consideración que el interés asegurable puede verse afectado por el nivel de importantica que tiene el tema de la unión familiar que ambas variables proporciona.

### 4.4. Aplicación Empírica de la Técnica de Rough Set

En esta sección, se analizan los resultados obtenidos de aplicar el enfoque Rough Set a la cartera muestra. En este caso, al igual que en el anterior, se busca es obtener un conjunto de patrones o reglas que resuman el conocimiento contenido en la base de datos y que sirva posteriormente para clasificar nuevas pólizas en función de las características de las mismas. A diferencia de los Árboles de Decisión, las reglas son sentencias lógicas y no se presentan en forma de árbol. En el presente estudio, se busca clasificar la cartera de pólizas hacia alguna de categorías: vigor o anulada con base en las variables cualitativas que describen a cada uno de los clientes de la muestra.

De la misma forma en cómo fueron utilizadas las variables para la aplicación de Árboles de Decisión, se tiene una muestra de 19.784 pólizas las cuales se han clasificado de acuerdo a las dos categorías: Vigor o Anulada con base en la variable TIPO PRESTACION. Esta variable toma el valor de 0 y 1 respectivamente *(Tabla 30)*:

| COD | TIPO_PRESTACION | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 0 | VIGOR | 16,568 | *83.74%* |
| 1 | ANULADA | 3,216 | *16.26%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 30:** *Distribución de acuerdo a la variable TIPO DE PRESTACION*
**Fuente**: *Propia de los autores*

Las variables de la muestra total han sido evaluadas de acuerdo a los valores que tomas las 14 variables cualitativas y cuantitativas definidas en el capítulo 3. Así mismo, la discretización de las variables no viene impuesta por la metodología Rough Set, pero facilita la aplicación e interpretación del modelo. Por lo que se considerará la discretización documentada en el tercer capítulo.

### 4.4.1. Resumen de Validación de Resultados bajo Rough Set

Uno de los primeros pasos que se debe realizar en la aplicación de este tipo de modelo, es la validación del mismo. Ahora bien, si se desarrolla un modelo y se validase con la misma muestra o con muestras que contienen las mismas observaciones, en este caso pólizas, aunque correspondan a diferentes años, los resultados podrían ponerse en cuestión. Para ello, se efectuado un procedimiento de *Validación Cruzada*[35], la cual consiste en hacer ciertas particiones de igual tamaño en los datos dejando una muestra para estimar el modelo y otro conjunto de datos para su validación. Este proceso se repite tantas veces como particiones se hacen. El resultado final es la media de los resultados obtenidos, con frecuencia se suele utilizar 10 particiones. Así se tiene que cuanto más más alta es la tasa de validación cruzada, mayor fiabilidad del modelo obtenido.

Pues bien, siendo así la aplicación de la metodología Rough Set ha presentado una precisión satisfactoria utilizando dicha validación cruzada en 10 pliegues. Dando como resultado un 83.7%, indicando un porcentaje de pólizas correctamente clasificadas considerablemente bueno elevando el poder de predicción del modelo *(Figura 35)*:

| | | 0 | 1 | No. of obj. | Accuracy | Coverage |
|---|---|---|---|---|---|---|
| Actual | 0 | 1,464.4 | 192.1 | 1,656.5 | 0.884 | 1 |
| | 1 | 130 | 191.5 | 321.5 | 0.595 | 1 |
| | True positive rate | 0.92 | 0.5 | | | |

Total number of tested objects: 1,978
Total accuracy: 0.837
Total coverage: 1

**Figura 35:** *Salida de Validación Cruzada (Rough Set)*
**Fuente**: *Propia de los autores*

[35] Para mayor claridad del concepto de Validación Cruzada, se puede consultar la explicación que ofrece Wikipedia en: *http://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada*. Mismo que también ha sido utilizado por Camacho Miñano, M. M. y Segovia Vargas, M. J. (2012)

Al igual que se hiciese en la aplicación de Arboles de Decisión, se observa la diagonal de la matriz de confusión. Se tiene que el valor de $a_{11}$=1.464,4 es mayor que $a_{21}$=191,5; y por otro lado, $a_{22}$=192,1 es mayor que $a_{12}$=130. Por lo que se tiene que un 59,5% de las pólizas anuladas son clasificadas correctamente y un 88,4% de las pólizas en vigor; dejando evidencia la fiabilidad del poder predictivo de las reglas de decisión obtenidas del modelo.

### 4.4.2. Resumen de las Principales Reglas

A continuación se presenta un extracto de los principales resultados que arroja el modelo. Esto es, una selección de las sentencias o reglas con mayor fuerza que reúnan el conjunto de variables que clasifican a la muestra de pólizas del estudio.

Previo a ello, un paso importante en la aplicación es la construcción del núcleo y los reductos. Los reductos se definen como el mínimo subconjunto de atributos independientes que aseguren la misma calidad de clasificación que la totalidad del conjunto de todos ellos (Laitinen (1992), García et al., (1997), McKee, (2000), Segovia (2003)). Ahora bien, la intersección de los reductos da como resultado al núcleo. En la aplicación, cabe mencionar que existe un único reducto elegido que contiene todas las variables, con excepción de la variable TIPO PRIMA. Y a su vez, como sólo existe un único reducto, éste coincide con el núcleo. Elegido el reducto, aquellas variables que no se encuentran dentro de éstos, pueden ser eliminadas de la base de datos; por lo que la única variable que se ha quitado de la información codificada es el TIPO PRIMA; ya que obviando esta información no se pierde capacidad de clasificación y el resto de variables son las mínimas para predicción de la anulación o no del contrato de seguros.

Ahora bien, para resumir el conjunto de reglas se han agrupado en dos cortes: el primero de ellos reuniendo las reglas que clasifican a la Clase 1 que corresponde a los patrones de comportamiento útiles para definir el perfil del cliente propenso a cancelar su póliza de seguros. Un segundo corte se muestra las reglas correspondientes a la Clase 2 siendo ésta la categoría que determinarían al tipo de

clientes susceptible a conservar su póliza y mantenerse en vigor dentro de la entidad. Dentro de esta segunda corte, dada la gran cantidad de reglas obtenidas, se han tomado las primeras 30 reglas con una fortaleza superior a 345, ya que con el análisis de estas reglas es suficiente para el estudio en cuestión.

#### 4.4.2.1. Reglas para la CLASE 1=Cancelación

Primeramente, se muestran el conjunto de reglas más fuertes que clasifican a la Categoría1=Cancelación de la Póliza *(Figura 36)*:

| | SEXO | EDAD ACTUARIAL | ANTIGUEDAD | TIPO PRODUCTO | FORMA PAGO | EDO CIVIL | HIJOS | VALOR CLIENTE | ICE | NIV INGRESOS | NIV ESTUDIOS | TIPO PRESTACION (clase) | FORTALEZA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | H | | 1 | 1 | 1 | 2 | | | | | 99 | 1 | 76 |
| 2 | H | | 1 | 1 | 1 | | N | | | | | 1 | 74 |
| 3 | | | 1 | 1 | 1 | | | A | | | | 1 | 71 |
| 4 | H | | 1 | 1 | 1 | | | | 4 | | | 1 | 47 |
| 5 | | 1 | 1 | 1 | 1 | 2 | N | | | | | 1 | 47 |
| 6 | | | 1 | 1 | 1 | 2 | N | | 4 | | | 1 | 46 |
| 7 | | 1 | 1 | 1 | 1 | 2 | | | | | 99 | 1 | 42 |
| 8 | H | 5 | 1 | | | | | A | | | | 1 | 39 |
| 9 | H | | 1 | 1 | | | | A | 4 | | | 1 | 39 |
| 10 | | 1 | 1 | 1 | | 2 | N | A | | | | 1 | 39 |
| 11 | | | 1 | 1 | 1 | 2 | N | | 3 | 9 | | 1 | 39 |
| 12 | | | 1 | 1 | 1 | 2 | N | | 3 | | 99 | 1 | 39 |
| 13 | | 5 | 1 | | | | | A | 3 | | | 1 | 38 |
| 14 | | 1 | 1 | 1 | | | N | A | | | 99 | 1 | 38 |
| 15 | H | 1 | 1 | 1 | 1 | | | | | | 99 | 1 | 38 |
| 16 | | 5 | 1 | 1 | | | N | A | | | | 1 | 37 |
| 17 | | 5 | 1 | 1 | 1 | 2 | | | | 9 | | 1 | 37 |
| 18 | | 5 | 1 | 1 | 1 | 2 | | | | | 99 | 1 | 37 |
| 19 | | 1 | 1 | 1 | 1 | 2 | | | 4 | | | 1 | 36 |
| 20 | | 1 | 1 | 1 | | 2 | | A | 4 | | | 1 | 36 |
| 21 | | 1 | 1 | 1 | | | | A | 4 | | 99 | 1 | 36 |
| 22 | | 1 | 1 | | 1 | 2 | N | A | | | | 1 | 35 |
| 23 | | 5 | 1 | | 1 | 2 | | | 3 | 9 | | 1 | 31 |
| 24 | | 5 | 1 | | 1 | 2 | | | 3 | | 99 | 1 | 31 |
| 25 | | 5 | 1 | | 1 | | N | | 3 | 9 | | 1 | 31 |
| 26 | | 5 | 1 | | 1 | | N | | 3 | | 99 | 1 | 31 |
| 27 | H | 5 | 1 | | 1 | 2 | | | | | | 1 | 30 |
| 28 | H | 5 | 1 | | 1 | | N | | | | | 1 | 30 |

**Figura 36**: *Reglas con Mayor Fuerza – Rough Set (CATEGORIA 1=Cancelación)*
**Fuente**: *Propia de los autores*

Cabe observar que no existe regla alguna donde aparezcan todos los atributos, lo que significa que no existe una combinación global de todos las características de un cliente que distingan exactamente el perfil del cliente "*cancelador*"; en otras palabras no existe alguna variable en especial que sobresalga sobre otras, sino por el contrario la interrelación entre los factores es lo que conforma cada una de las opciones que llevarían a definir y categorizar a los clientes. Sin embargo, si existe la posibilidad de

que dadas ciertas condiciones o tipos de clientes determinan nivel de riesgo de anulación o caída de cartera.

#### 4.4.2.2. Reglas para la CLASE 0=Retención

Por otro lado, se tiene el conjunto de reglas que clasifican a la Categoría 0=Retención de la Póliza (*Figura 37):*

| | SEXO | EDAD ACTUARIAL | ANTIGUEDAD | TIPO PRODUCTO | FORMA PAGO | EDO CIVIL | HIJOS | VALOR CLIENTE | ICE | NIV INGRESOS | NIV ESTUDIOS | TIPO PRESTACION (clase) | FORTALEZA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | 8 | 1 | | | | | | | | 0 | 1060 |
| 2 | V | | 6 | 1 | | | | | | | 99 | 0 | 971 |
| 3 | | | 6 | 1 | 5 | | | | | | | 0 | 805 |
| 4 | | | 7 | 1 | | | N | | | | 99 | 0 | 728 |
| 5 | | | 7 | 1 | | | N | | | 9 | | 0 | 709 |
| 6 | V | | 6 | 1 | | | N | | | | | 0 | 702 |
| 7 | | | 8 | | 6 | | | | | | | 0 | 660 |
| 8 | | | 6 | 1 | | | | A | | 9 | | 0 | 649 |
| 9 | | | 6 | 1 | | | | | 4 | | 99 | 0 | 643 |
| 10 | | | 6 | 1 | | | S | | | 9 | 99 | 0 | 640 |
| 11 | | 5 | 6 | 1 | | | | | | | | 0 | 638 |
| 12 | | | 7 | 1 | | 2 | N | | | | | 0 | 630 |
| 13 | | | 6 | 1 | | | | | 3 | 9 | | 0 | 540 |
| 14 | H | | 7 | 1 | | | | | | | | 0 | 532 |
| 15 | | | 7 | 1 | | | | A | | | | 0 | 532 |
| 16 | V | | 6 | | 6 | | | | | | 99 | 0 | 524 |
| 17 | | | 7 | 1 | 5 | | | | | | | 0 | 492 |
| 18 | | 5 | 6 | | 6 | | | | | | | 0 | 487 |
| 19 | | | 10 | | | | | | | | | 0 | 447 |
| 20 | | | 6 | 1 | | | N | A | | | | 0 | 418 |
| 21 | V | | 6 | 1 | | | | A | | | | 0 | 417 |
| 22 | | | 6 | 1 | | | N | | 4 | | | 0 | 412 |
| 23 | | | 6 | | 6 | | | A | | 9 | | 0 | 392 |
| 24 | | | 6 | 1 | | | N | | 3 | | | 0 | 390 |
| 25 | H | | 6 | | 5 | | | | | | | 0 | 362 |
| 26 | V | | 6 | | 6 | | N | | | | | 0 | 361 |
| 27 | | | 6 | | 6 | | S | | | 9 | 99 | 0 | 359 |
| 28 | | | 6 | | 6 | | | | 4 | | 99 | 0 | 349 |
| 29 | | | 7 | 1 | | | | | 3 | | 99 | 0 | 348 |
| 30 | | | 7 | | 6 | | N | | | | 99 | 0 | 345 |

**Figura 37***: Reglas con Mayor Fuerza – Rough Set (CATEGORIA 0=Retención)*
**Fuente**: *Propia de los autores*

Esta categoría, que corresponde a los clientes susceptibles a conservar sus pólizas con la entidad aseguradora, se clasifican mejor, no sólo por la mayor fuerza que presentan sus reglas, sino por la cantidad de reglas de decisión que resultan. De esta forma, aunque no es el objetivo principal del estudio, obtener patrones que caractericen a los clientes propensos a no cancelar su contrato de seguros; pueden ser resultados de gran utilidad para las compañías.

Así se puede decir que es más complicado obtener reglas globales para esta clase, es decir pueden existir varias características que definirían al perfil de los clientes propensos a conservar su contrato de seguros.

### 4.4.3. Resumen de las Principales Variables

Analizando el conjunto de reglas presentadas con anterioridad, se puede observar que no existen reglas universales que definan al tipo de cliente propenso a anular su póliza de seguros. Algo similar ocurre con el evento contrario, es decir, con las características precisas que determinan al conjunto de clientes que conservan sus contratos.

Sin embargo, se pueden detectar las variables que aparecen más frecuentemente y dentro de las reglas con mayor fuerza. De esta forma, lograr obtener un conjunto mínimo de variables a partir de la muestra, que aseguran la mejor calidad de clasificación del total de variables utilizadas en el modelo.

Primeramente para la Categoría 1=Cancelación, se tiene que la longitud de las reglas cuenta con 2 variables como mínimo; y como máximo 8 variables. Las variables que se presentan con mayor frecuencia son: ANTIGÜEDAD, FORMA DE PAGO y TIPO PRODUCTO; sin dejar de lado las variables ESTADO CIVIL e HIJOS que también aparecen dentro de las principales reglas presentadas anteriormente.

Por el contrario, para la Categoría 0=Retención se cuenta con reglas cuya longitud van desde 2 variables con un máximo de 6 variables. Nuevamente, las variables que se presentan con mayor frecuencia son: ANTIGÜEDAD, TIPO PRODUCTO; sin embargo, en esta Clase también se obtiene la FORMA DE PAGO con menor frecuencia y al mismo nivel de aparición que la variable HIJOS. Dentro de esta categoría no destaca la variable ESTADO CIVIL como un patrón de comportamiento para definir al cliente que conserva su póliza de seguros.

A manera de resumen general de las variables más significativas para cada una de las categorías, se tiene (*Tabla 31*):

**ROUGH SET**

**RESUMEN REGLAS**

| Categoría: 1 | | Categoría: 0 | |
|---|---|---|---|
| CAIDA | | RETENCION | |
| | | | |
| VARIABLES SIGNIFICATIVAS | | VARIABLES SIGNIFICATIVAS | |
| * | Antigüedad | * | Antigüedad |
| * | Forma Pago | * | Tipo Producto |
| * | Tipo Producto | * | Forma Pago |
| * | Edo Civil | | |
| * | Hijos | | |

**Tabla 31**: *Resumen de Resultados Rough Set – AMBAS CATEGORIAS: 1-CAIDA y 0-RETENCION*
**Fuente**: *Propia de los autores*

Así finalmente, se puede concluir que las variables que en ambas categorías están contenidas con mayor frecuencia y por tanto, pueden ser consideradas como los atributos que clasificarían el éxito o fracaso de la conservación o anulación de la cartera de pólizas de una compañía aseguradora son: ANTIGÜEDAD, TIPO PRODUCTO y FORMA PAGO.

### 4.4.4. Principales Resultados Obtenidos bajo Rough Set

Así se tiene que las variables con mayor frecuencia que pudiesen sugerir los posibles patrones de comportamiento en los clientes para evaluar si posible abandono o conservación de su contrato de seguros serían:

- La duración o ANTIGÜEDAD que tiene el cliente dentro de la compañía de seguros podría ser una de los temas más sensibles para la elección de cancelación de un asegurado. Esto, puede venir motivado a que existen ciertos derechos ganados por el cliente por el hecho de mantener vigente de manera continua e ininterrumpida su póliza de seguros.

Por otro lado, también se debe tener en cuenta el tema de fidelización de los clientes hacia la marca. Es decir, pueden existir estrategias de retención de clientes,

donde se busca conservar y mantener “fiel” al cliente con el producto contrato; de esta forma, se fortalece la relación cliente-aseguradora y con ello de manera indirecta se mitiga el riesgo de abandono que pudiese afectar en la rentabilidad de la cartera de pólizas.

- Otra de las variables resultante es el TIPO PRODUCTO contrato por el cliente. Esto habla de tener en cuenta que el riesgo de Caída de Cartera dependerá del tipo de cartera o *mix de negocio* que tiene la entidad aseguradora.

Más que ser un factor determinante en un cliente para su elección en cancelar o no su póliza de seguros; se puede ver como una variable para el seguimiento del riesgo. Esto es, que con base en este resultado, se podría sugerir que exista alguna herramienta de alarmas para cierto tipo de productos con mayor propensión a su anulación. De esta forma, poder tener estrategias de negocio enfocadas al seguimiento de algún determinado producto donde se pudiese presentar con mayor intensidad este riesgo. Y con base en ello, poder mantener el *mix negocio* adecuado que mantenga la suficiencia del volumen de primas y por ende, asegurar la rentabilidad del negocio en su globalidad.

- Por otro lado, la FORMA PAGO es una tercera variable a ser considerada como factor determinante en la elección de cancelación de la póliza contratada. Esto representa la posibilidad que existe de que el cliente se plantee continuamente el hecho de renovar, mantener y pagar la prima de su seguro. En otras palabras, en cuanto más existe la posibilidad de que un cliente se cuestione este hecho, más tendrá la oportunidad de cancelar su póliza. Es decir, un posible que un cliente cuya Forma de Pago de su póliza de seguros es Anual o Semestral, se cuestione menos el hecho de continuar o no con su póliza que aquel cliente que cuenta con una Forma de Pago Mensual; ya que éste último, durante 12 veces al año, se plantea la idea de mantener vigente su contrato.

Ahora bien, una vez más, esta variable además de aportar un patrón de comportamiento en el tipo de clientes propensos a la cancelación de su póliza; sugiere un factor a ser incluido dentro de los controles y alarmas que deba implementar la entidad aseguradora para gestionar el riesgo de caída de cartera. O bien, un factor a ser considerado en las estrategias de gestión de la cartera de clientes, así como en los análisis de clientes rentables y grupos de riesgo preferentes que desee conservar la entidad. Todo ello, con el fin de una gestión eficiente de la retención de clientes y volumen de primas; que se traduce en una gestión del riesgo óptima.

## CAPITULO 5: APLICACIÓN DE LA METODOLOGÍA DE MODELOS LINEALES GENRALIZADOS

### 5.1. Introducción

Dentro de la ciencia actuarial, a la salida de una persona de un determinado grupo se le conceptualiza como *decremento*. Es así como los cálculos actuariales se centran en el cálculo de las probabilidades de permanecer o salir de cierto grupo por una serie de causas o *decrementos* como son: muerte, invalidez, rescate, anulación, etc. De aquí que uno de los principales objetivos de las entidades aseguradoras contempla el conocimiento, cálculo, análisis y gestión de dichos *decrementos* transformados en términos de tasas del riesgo de mortalidad, incapacidad, caídas y rescates.

Ahora bien, con la nueva regulación propuesta por Solvencia II, las compañías aseguradoras están siendo sometidas a desarrollar nuevas técnicas para la cuantificación y control de los riesgos a los que se encuentran expuestas. Todo ello con el fin de lograr implementar una gestión integral del riesgo que contemple un adecuado nivel de solvencia. Dicha gestión de riesgos implica contemplar todos y cada uno de los componentes del negocio asegurador que puedan generar algún tipo de contingencia para la compañía.

Por un lado, el proyecto de Solvencia II[36] propone que exista una evaluación constante de la precisión de los cálculos realizados por las entidades aseguradoras, a lo cual identifica como "*best-estimate*". Por otro lado, las entidades aseguradoras derivan sus cálculos de *decrementos* basándose, generalmente, en su experiencia histórica asumiendo que el pasado sería un buen indicador de lo que ocurrirá en el futuro. Uniendo estos dos conceptos, la nueva regulación promueve la importancia de un adecuado análisis de sus riesgos mediante el cálculo del "*mejor-estimador*" de las

[36]A través de los Principios del Market-Consistent Embedded Value *(MCEV Principales*, CFO Forum, June 2008.

causas o *decrementos* que puedan generar algún tipo de contingencia para la compañía.

En otras palabras, procurar utilizar parámetros o hipótesis específicas y prudentes que reflejen los riegos reales a los que está expuesta una entidad aseguradora, de la "mejor" u óptima manera de estimarlo. Uno de dichos riesgos contemplados es la caída de cartera que registra una entidad entendiéndose como tal a la rotación o salida de asegurados, lo cual se ve directamente reflejado en el decrecimiento en el volumen de primas de la entidad (Millán Aguilar, Adolfo et. al. 2000). Generalmente, el riesgo de caída de cartera ha sido calculado considerando que dicho evento se encuentra relacionado con el tipo de producto, tiempo de duración en que ha estado vigente la póliza, o bien el año de emisión de la póliza. Sin embargo, dicho riesgo se puede ver inducido por otra serie de factores subyacentes o tendencias del tipo de clientes que contratan un seguro (edad, género, geografía).

Existen diversas metodologías utilizadas para la estimación de las anulaciones que se producirán en el futuro y harían fluctuar el volumen del negocio y márgenes de rentabilidad; que se traduce en la probabilidad de cancelación del contrato de seguros basado en la experiencia registrada en años anteriores. La mayor parte de dichas metodologías recurren a técnicas estadísticas que, mediante un coeficiente de caída, recogen el promedio de porcentajes de caída registrados durante el histórico de la cartera.

Sin embargo, la utilización de dichas técnicas muestra poco margen de maniobra en cuanto a la gestión del riesgo como tal; ya que la visión puramente matemática que proporcionan estas metodologías, niegan la posibilidad de la inclusión de componentes cualitativos que maticen el resultado de tal forma que se pueda incurrir en él. En otras palabras, mediante una adecuada definición del apetito de riesgo que pretenda una entidad aseguradora y el estudio de una serie de factores cualitativos que incurren en la decisión de permanencia o abandono en un cliente, se puede lograr una gestión y control del riesgo de caída de cartera mucho más manipulable y alineada con la estrategia de negocio planteada por la entidad.

Ahora bien, los Modelos Lineales Generalizados (*GLM –Generalized Linear Models* –) introducidos a comienzos de los años 70 (Nelder y Wedderbum, 1972), se han convertido en una de las principales herramientas de análisis estadístico en toda clase de áreas. No fue hasta los años 90, cuando se comenzaron a utilizar dentro de la Estadística Actuarial como una herramienta utilizada para temas de tarificación dentro del sector asegurador (Guillén Estany et al., 2005; Ohlsson y Johansson, 2010). De acuerdo con esta metodología las primas son calculadas tras un análisis de regresión en el que se obtiene como variable respuesta o dependiente (número de siniestros o importe reclamado), basándose en un conjunto de variables explicativas, es decir, una serie de factores relacionados con el evento que simula dicha variable respuesta (generalmente características propias del asegurado de la póliza).

Sin embargo, existen pocos estudios que utilizan la metodología ofrecida por los GLM para el análisis del riesgo de caída de cartera al que está expuesta una entidad aseguradora (Cerchiara, R.R. et. al. 2008). Aun cuando el objetivo de este estudio no es llegar al cálculo exacto y robusto de la caída de cartera por medio de un modelo GLM, sí que es posible investigar y proponer el uso de esta metodología para la calibración de este riesgo. Es decir, con el fin de conocer y entender los componentes que puedan estar relacionados con la propensión, que tienen los asegurados, en la cancelación de su póliza; lograr gestionar las causas y factores que inciden en el riesgo de caída de cartera.

Siendo así, el objetivo del presente estudio es, mediante la utilización empírica de dicha metodología, lograr identificar la información o características del asegurado que describan el tipo de clientes propensos a la anulación de su contrato de seguros. Así mismo, la metodología de los GLM podría ofrecer una herramienta que reconozca ciertas relaciones no lineales que podrían ayudar al análisis de los parámetros que afectan a este riesgo; y de esta forma, permitir tener conocimiento sobre las correlaciones y dependencias de los factores que lo propician con el fin de lograr un control y gestión del riesgo en su globalidad.

Así se tiene que en el contexto actual del mercado asegurador en donde existe una disminución del volumen de negocio y creciente tendencia de pérdida de la

cartera; cobra importancia el tema de poder implementar estrategias para la retención de clientes y lograr orientar la toma de decisiones por medio de la localización de grupos de riesgos, definidos por ciertas características concretas y medibles del tipo de cliente "*cancelador*".

De aquí el objetivo del presente estudio, el cual se encuentra estructurado de la siguiente forma. En la siguiente sección, se revisará el marco teórico sobre el que descansan los Modelos Lineales Generalizados; así como sus principales características, estructura y componentes. Así mismo, se describirán la serie de fases que se deben seguir para la aplicación de un modelo GLM, así como una recopilación de las aplicaciones de estos modelos que se han propuesto dentro del sector asegurador de vida, siguiendo a como han tratado esta metodología algunos autores (Heller, Gillian et. al. 2008, Lindsey, James 1997 y Nelder, J.A. et. al. 1989).

Dentro de la tercera sección, se realizará una aplicación práctica de la metodología que ofrecen los GLM a una cartera real de clientes de una compañía de seguros centrado en el ramo de Vida Individual; procurando seguir las fases de procesos y análisis recomendados para lograr obtener la mejor aplicación empírica posible.

En la sección 4, se llegará a la identificación de las características o variables explicativas que puedan describir al tipo de cliente susceptible a la cancelación de su póliza, mediante los resultados obtenidos de la aplicación práctica.

De esta forma, se finalizará con una quinta sección enfocada a las conclusiones y futuras líneas de investigación, así como limitaciones y principales contribuciones del presente estudio; que si bien, busca dar un enfoque distinto, a su vez, ofrezca una metodología que ayude a contrastar los resultados obtenidos con otros modelos sugeridos.

### 5.2. Modelos Lineales Generalizados

Tradicionalmente las compañías aseguradoras recurren a métodos estadísticos para cuantificar el riesgo de caídas y rescates. Lo hacen midiendo la tasa de anulación mediante el cálculo de número de pólizas anuladas y/o rescatadas entre una cierta exposición considerada; incluyendo en este análisis de siniestros una determinada segmentación: duración, tipo de producto (Guillen, et. al. 2008).

Sin embargo, esta metodología presenta algunos inconvenientes. Por un lado, no deja de ser un método puramente univariante donde la estadística se centra en una única característica o variable de manera excluyente. De ahí otro de los puntos débiles de esta metodología tradicional, ya que se centra en unos pocos y limitados factores de riesgo; mientras que la caída de cartera se puede ver influenciada por un gran número de factores, actuando sobre un determinado evento de forma simultánea., y no sólo en dos o tres.

Es así como se encuentran los modelos predictivos como una herramienta potente que ayudan a resolver estos inconvenientes y restricciones que presenta el modelo tradicional; ofreciendo una alternativa para analizar la caída de cartera permitiendo mayor capacidad de interacción entre los factores de riesgo, y a la vez facilita analizar el verdadero impacto de cada factor.

Los modelos predictivos trabajan relacionando un evento determinado (en este caso, la anulación o caída de cartera que puede presentar una entidad aseguradora); con un cierto número de factores *(Figura 38)*:

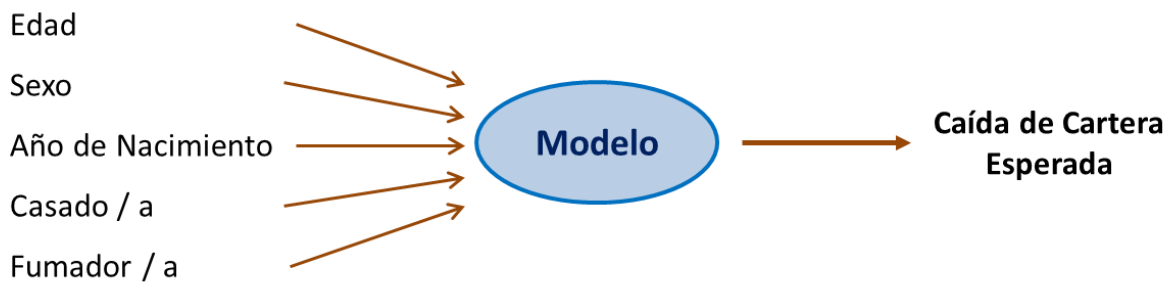


**Figura 38:** Traducción de Modelo Predictivo en Seguros de Vida
**Fuente**: *Propia de los autores*

Dentro de dichos modelos predictivos, se encuentran los Modelos Lineales Generalizados (*GLM*, por sus siglas en inglés), que constituyen una generalización de los tradicionales Modelos Lineales (*LM – Linear Models* –); donde se asume que el valor esperado de la variable dependiente se encuentra condicionado a las variables independientes expresándose como una combinación lineal de los valores que dichas variables. Pues bien, en el caso delos GLM, se trata de un método capaz de modelizar un número como función de varios factores.

Dentro del sector asegurador, esto se traduciría en la construcción de un modelo de costes de siniestros (anulaciones o cancelaciones), permitiendo observar la influencia de varios factores de riesgo; esto es, mostrando una capacidad para tener en cuenta automáticamente las correlaciones que existe entre los datos para la estimación de dichos costes de siniestralidad. En otras palabras, se podría demostrar cómo ciertos parámetros tradicionalmente ignorados (como por ejemplo el nivel socio-económico o la forma de pago) pueden afectar al comportamiento de los asegurados; es decir, lograr investigar la realidad subyacente de los parámetros que podrían brindar los datos utilizados para el análisis.

### 5.2.1. Marco Teórico

Los Modelos Lineales Generalizados (*Generalized Linear Models,* en inglés - *GLM* -), introducidos a comienzos de los años 70 (Nelder y Wedderburn, 1972), resumen un grupo homogéneo de métodos de regresión (logística, Poisson, gamma, etc.), previamente consideradas de forma independiente. La amplia difusión que han tenido, los ha convertido en una de las principales herramientas de análisis estadístico en toda clase de áreas.

Los GLM se utilizan para cuantificar la relación entre a variable $Y$, conocida como variable respuesta, a través de $X$ variables explicativas. Dicho de esta forma burda, esto concuerda con el concepto de los Modelos Lineales Simples; lo cual es lógico ya que los GLM constituyen una extensión de los clásicos modelos lineales; compartiendo el mismo punto de partida. Sin embargo, los GLM cuentan con su propia estructura, elementos y método de análisis e interpretación de los resultados, que los hace más atractivos y ofrece mayor facilidad en su aplicación.

Siendo así, el objetivo del presente estudio es tratar de aclarar si este tipo de modelos son susceptibles para la cuantificación del riesgo de caída de cartera. Pero antes de proceder a su aplicación, es necesario primero englobar su metodología (De Jong, P.y Heller, G. Z. 2008); así como comprender sus requisitos y en términos generales conocer en que consiste su técnica para determinar si es aplicable a nuestro estudio.

### 5.2.2. Estructura y Parámetros

La estructura de un GLM presenta una relación lineal entre las variables explicativas y una transformación de la media de la variable respuesta. Esto es, que no existe una relación líneas entre ambas, sino entre una función de enlace (*función "link"*) y las variables explicativas:

$$g(E(Y)) = \sum_i \beta_i x_i$$

Ahora bien, en cuanto a sus parámetros, se tiene:

**MEDIA:**

$$\mu = E[Y_i] = g^{-1}\left(\sum_j \beta_j X_{ij} + \xi_i\right)$$

**VARIANZA:**

$$Var[Y_i] = \frac{\phi V(\mu_i)}{\omega_i}$$

Donde:

$Y_i$ =Vector de la Variable Respuesta

$g(x)$ =Función de Enlace

$X_{ij}$ =Matriz de factores

$\beta_j$ =Vector de parámetros del modelo

$\xi_i$ =Vector de efectos conocidos (*offset*)

$\phi$ =Parámetro escalar de la función $V(x)$

$V(x)$ =Función Varianza

$\omega_i$ =Peso asignado a cada observación (*prior weight*)

Una de las condiciones de los Modelo Lineales es el hecho de que exigen que la variable dependiente $Y$, condicionada a los valores de las $x_i$, siguen una distribución de probabilidad normal. Sin embargo, si dicha variable respuesta es discreta, entonces el modelo no funciona. Es aquí donde los GLM hacen presencia, ya que permiten modelar variables respuesta, ya sean continuas o categóricas.

Es así como una de las diferencias con respecto a los modelos lineales, es que en un GLM, la variable respuesta no tiene por qué seguir una distribución normal. En

otras palabras, los GLM unifican los modelos con variables de respuesta continua y categórica dando la posibilidad de analizar variables con distribuciones pertenecientes a la Familia Exponencial. Esto ayuda a reflejar los supuestos sobre la distribución de las variaciones inexplicadas, incluyendo no sólo a la Normal, sino a muchas otras de las más usadas en las aplicaciones, como son las distribuciones Binomial, Poisson, Gamma, etc.

Otra gran diferencia, es que en un GLM se distinguen 3 elementos (que a diferencia de un modelo lineal se tienen sólo dos elementos: variable a explicar y las variables explicativas):

- La variable respuesta (con $n$ observaciones y todas ellas con la misma distribución)

- El conjunto de variables explicativas (con sus correspondientes parámetros)

- Una función de enlace (que relaciona en linealidad entre la variable respuesta y las variables explicativas)

### 5.2.3. Componentes

Como ya se ha comentado, los modelos lineales generalizados son una extensión de los modelos lineales clásicos, por lo que comparten algunos de sus componentes. Siendo así, los Modelos Lineales Generalizados tienen tres componentes básicos, que se detallan a continuación:

- **Componente Aleatorio:**

Identifica la variable respuesta y su distribución de probabilidad. Este componente consiste en una variable aleatoria $Y$ con observaciones independientes $(y_1, \dots, y_N)$.

En muchas aplicaciones las observaciones de $Y$ son binarias y se identifican como éxito y fracaso. Aunque de modo más general, cada $Y_i$ indicaría el número de éxitos de entre un número fijo de ensayos; y se modelarizaría como una distribución binomial. En otras ocasiones cada observación es un recuento, con lo que se puede asignar a $Y$ una distribución de Poisson o una distribución binomial negativa.

Finalmente, si las observaciones son continuas se puede asumir para $Y$ una distribución normal. Todos estos modelos se pueden incluir dentro de la llamada familia exponencial de distribuciones.

- **Componente Sistemático:**

Específica las variables explicativas, que entran en forma de efectos fijos en un modelo lineal, es decir las variables $x_j$que se relacionan como:

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

Esta combinación lineal de variables explicativas se denomina *predictor lineal*, el cual se puede expresar como:

$$\eta = \sum_j \beta_j x_j$$

- **Función Link:**

Se denota el valor esperado de $Y$ como $\mu = E[Y]$, entonces la *función link* especifica una función que relaciona a $\mu$ con el predictor lineal como:

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

En otras palabras, la función link $g(\cdot)$ o función enlace que relaciona el componente aleatorio y el componente sistemático; de tal forma que:

$$\eta = g(\mu) = \sum_j \beta_j x_j$$

Si la función enlace se supone más simple $g(\mu) = \mu$, es decir la identidad, esto daría lugar al modelo de regresión lineal clásico, esto es:

$$\mu = E[Y] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

Los modelos de regresión lineal típicos para respuestas continuas son un caso particular de los GLM. Estos modelos generalizan la regresión ordinaria de dos modos: Permitiendo que $Y$ tenga distribuciones diferentes a la normal; y por otro lado, incluyendo distintas funciones link de la media. Esto resulta bastante útil para datos categóricos.

### 5.2.4. Familia Exponencial

La Familia Exponencial de distribuciones es uno de los conceptos clave en los Modelos Lineales Generalizados, ya que garantiza la equivalencia entre los métodos de máxima verosimilitud y el método de mínimos cuadrados ponderados, para estimar los parámetros desconocidos del modelo.

Las funciones de probabilidad dentro de esta familia, se pueden expresar de la siguiente forma general:

$$f(y) = c(y, \phi) exp\left\{\frac{y\theta - a(\theta)}{\phi}\right\}$$

Siendo $\theta$ el parámetro canónico y $\phi$ el parámetro de dispersión; y de tal forma que la elección de las funciones $a(\theta)$ y $c(y, \phi)$ determinan la función de probabilidad según una distribución normal, binomial o gamma.

Por otro lado, se tiene que dos de las propiedades que comparten las distribuciones de la familia exponencial son:

- La distribución es especificada en términos de sus media y varianza
- La varianza de $Y$ es una función de su media

$$Var(Y) = \frac{\phi V(\mu)}{\omega}$$

Siendo así, las distribuciones pertenecientes a la Familia Exponencial con sus correspondientes parámetros canónicos y función varianza, resumidos en la siguiente tabla *(Tabla 32)*:

| DISTRIBUCION | Notación | $a(\theta)$ | $\phi$ | $V(\mu)$ |
|---|---|---|---|---|
| ***Normal*** | $N(\mu, \sigma^2)$ | $\theta$ | $\sigma^2$ | $1$ |
| ***Poisson*** | $P(\mu)$ | $e^{\theta}$ | $1$ | $\mu$ |
| ***Gamma*** | $G(\mu, v)$ | $-1/\theta$ | $v^{-1}$ | $\mu^2$ |
| ***Binomial*** | $B(m, \pi)/m$ | $e^{\theta}/(1+e^{\theta})$ | $1/\mathrm{m}$ | $\mu(1-\mu)$ |
| ***Inversa Gaussiana*** | $IG(\mu, \sigma^2/\omega)$ | $(-2\theta)^{-1/2}$ | $\sigma^2$ | $\mu^3$ |

**Tabla 32:** Distribuciones de la Familia Exponencial (parámetros y función de varianza)
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

### 5.2.5. Función Enlace

Como ya se ha comentado, existe una relación entre la variable respuesta y las explicativas; la cual no siempre corresponde a una relación lineal entre ambas. Pues bien, de aquí surge el concepto de *"función vínculo"* o *"función enlace"*; quien se ocupa de linealizar la relación entre la variable dependiente y las variables explicativas mediante la transformación de la variable respuesta.

En términos técnicos, la función enlace relaciona al predictor lineal $\eta$ con el valor esperado de $\mu$. En otras palabras, es más útil considerar a $\mu$ como una función del predictor lineal, es decir la inversa de $g(x)$ es considerada como:

$$\mu = g^{-1}(\eta)$$

En los modelos lineales clásicos, la media y el predictor lineal son idénticos; por lo que la función vínculo es la identidad. Por otro lado, la función enlace debe satisfacer la condición de ser una función monótona y diferenciable.

En la siguiente tabla, se resumen algunas de las funciones vínculos comúnmente más utilizados *(Tabla 33)*:

| FUNCION VINCULO | $g(x)$ | $g^{-1}(x)$ |
|---|---|---|
| ***Identidad*** | $x$ | $x$ |
| ***Log*** | $\ln(x)$ | $e^x$ |
| ***Logit*** | $\ln(x/(1-x))$ | $e^x/(1+e^x))$ |
| ***Reciprocal*** | $1/x$ | $1/x$ |

**Tabla 33:** Funciones Vinculo
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Cuando se aplica cierta función vínculo no significa que sea la única o que siempre sea la más adecuada para el caso de estudio. Es por ello, que es recomendable seleccionar más de una función enlace para el mismo modelo y observar con cuál se obtiene un mejor ajuste del modelo a los datos estudiados.

### 5.2.6. Offset

Ahora bien, retomando la función enlace genérica definida como:

$$g(\mu) = \eta = x\beta$$

en ocasiones, para que se pueda estimar dicha expresión, se debe incluir en la ecuación lo que se llama *offset* o *añadido*; lo cual se trata de una constante que sirve para equilibrar el modelo. Dependiendo de la estructura de los datos, se estima y si resulta significativo, entonces se introduce como constante.

Una de las aplicaciones de este tipo de modelos, se trata de estimar el número de siniestros o las muertes dentro de un grupo de riesgo. Pues bien, en ocasiones es conocido el efecto que puede tener cierta variable al momento de modelar el conteo de siniestros; y es válido incluir dicha información en el modelo. Por ejemplo,

introducir el efecto de la exposición o número de expuestos al riesgo como una especie de "corrección" o ponderación de las observaciones de la muestra.

Esto puede lograrse introduciendo el término de offset dentro de la definición del predictor lineal $\eta$; quedando de la siguiente forma:

$$\eta = x\beta + \xi$$

lo que se traduce en:

$$E[Y] = \mu = g^{-1}(\eta) = g^{-1}(x\beta + \xi)$$

El offset es efectivamente otra variable $x$ de la regresión, con un coeficiente $\beta = 1$. De esta forma, $y$ obtiene el valor esperado directamente proporcional a la exposición. En otras palabras dicho, los términos *offset* se utilizan para corregir el tamaño de la muestra o diferir los períodos de observación.

### 5.2.7. Estimación

Una vez definido la estructura del modelo a seguir, los estimadores $\beta$ se ajustan a partir de la muestra de observaciones de $Y = (y_1, \dots, y_N)$; esto es, los parámetros se estiman a partir de la propia muestra. Pues bien, dicha estimación de parámetros puede hacerse a partir de varios métodos conocidos:

- Método de los Momentos

Es un método intuitivo de estimación de parámetros de una ecuación de regresión. Consiste en tomar como estimadores de los momentos de la población a los momentos de la muestra. En términos generales, se trata de resolver el sistema de equivalencias entre unos adecuados *momentos empíricos (muestrales)* y *teóricos (poblacionales);* es decir, que la media de la población y la varianza son iguales a sus equivalentes de la muestra.

- Estimadores Máximo-Verosimilitud

Se trata de otro método habitual para ajustar un modelo y encontrar sus parámetros. La verosimilitud consiste en otorgar a un estimador una determinada "credibilidad" a cierto valor (estimador). En términos probabilísticos se puede decir que la verosimilitud es la probabilidad de que ocurra una determinada muestra, si es cierta la estimación que se ha efectuado o el estimador que se ha planteado. Por lo que la máxima verosimilitud será aquel estimador que nos arroja mayor *"credibilidad"*. El ´método de máxima verosimilitud elige los valores de los parámetros que maximizan la probabilidad de haber observado la muestra $Y = (y_1, \dots, y_N)$

- Mínimos Cuadrados Ordinarios

Es otro método para encontrar los parámetros poblacionales en un modelo de regresión. En este caso, el método minimiza la suma de las distancias verticales entre las respuestas observadas en la muestra y las respuestas del modelo. Este método será consistente siempre y cuando no exista multicolinealidad y no haya autocorrelación. En estas condiciones, el modelo proporciona un estimador insesgado de varianza mínima siempre que los errores tengan varianzas finitas.

### 5.2.8. Estructuras de Modelos Comunes

Dentro del mercado asegurador, algunas de las situaciones más comunes por modelar son las frecuencias de cierto evento; o bien el importe de siniestro medio reclamado conocido como severidad de la siniestralidad ocurrida.

Para el caso de la estimación de frecuencias de siniestros, generalmente se asume que el uso de una distribución *Poisson* es lo más apropiado (Duncan Anderson, et. al. 2007). Siendo así, se suele utilizar como "peso" o ponderación de las observaciones, el nivel de exposición de cada una de éstas (es decir, el *offset* utilizado sería el *log* del nivel de *exposición*). Por otro lado, el modelo más común para estimar la severidad o costo medio de las reclamaciones es mediante el uso de la distribución *Gamma*.

Por lo tanto, se puede decir que, dependiendo de la naturaleza de los valores que toma la variable $Y = (y_1, \dots, y_N)$; se deberá tomar la combinación idónea de la distribución de probabilidad y la función de enlace que mejor se adapte al objetivo que se plantea. De esta forma, algunos de los modelos más comunes con sus respectivas estructuras, se resumen a continuación *(Tabla 34)*:

| $Y$ | Frecuencias de Siniestros | Número de Siniestros | Coste Medio de Siniestros | Probabilidad (renovación / cancelación) |
|---|---|---|---|---|
| **Función Enlace** $g(x)$ | $\ln(x)$ | $\ln(x)$ | $\ln(x)$ | $\ln(x/(1-x))$ |
| **Error** | Poisson | Poisson | Gamma | Binomial |
| **Parámetro Escalar** | 1 | 1 | Estimado | 1 |
| **Función Varianza** $V(x)$ | $x$ | $x$ | $x^2$ | $x(1-x)$ |
| **Pesos** $\omega$ | Exposición | 1 | # siniestros | 1 |
| **Offset** $\xi$ | 0 | ln(Exposición) | 0 | 0 |

**Tabla 34:** Estructuras de Modelos más comumes
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

### 5.2.9. Validación del Modelo

Una de las cuestiones principales que surge durante la aplicación de un modelo es la discrepancia o desviación que existe entre éste y las observaciones de la muestra tratada; de ahí la necesidad de considerar un proceso de ajuste o validación del modelo estadístico. Los principios de pruebas de significancia, el modelo de selección y las pruebas de diagnóstico son similares tanto en los GLM como para los modelos clásicos de regresión; aunque ciertos detalles técnicos difieren un poco entre ambos.

En cualquier proceso de ajuste de un modelo de datos se puede considerar como una forma de reemplazar un conjunto de datos $Y = (y_1, \dots, y_N)$ por ciertos valores ajustados $\mu$ que surgen de un modelo que implica un número determinado conjunto de parámetros. En términos generales, los valores de $\mu$ no serán exactamente iguales a los datos de $Y$; por lo que surge la necesidad de saber que tanto varían los valores entre ellos.

El modelo más simple, conocido también como *modelo nulo*, es el que tiene un solo parámetro que representa una $\mu$ común para todo el conjunto de $Y = (y_1, \dots, y_N)$ y asume toda la variación entre éstas con el componente aleatorio. Por el contrario, un *modelo saturado* es aquel modelo completo que cuenta con $n$ parámetros, siendo un parámetro para cada observación donde los valores de $\mu$ que derivan del modelo encajan perfectamente con los datos; sin embargo, asume todas las variaciones entre las $y$ al componente sistemático, sin dejar nada para el componente aleatorio.

En la práctica, un *modelo nulo* suele ser demasiado simple y un *modelo saturado* no resulta operativo ya que no resume los datos sino que simplemente se repiten en su totalidad. Sin embargo, éste último es útil como base para medir la diferencia o discrepancia que existe comparado con un modelo "intermedio" con $\mu$ parámetros.

De aquí la lógica que siguen las distintas técnicas de validación de un modelo de datos:

- Análisis de la Devianza

La Devianza se define como una medida de distancia entre los modelos saturados y ajustados. Siendo así, arroja una medida de bondad de ajuste entre los datos observados y los valores ajustados que se obtienen del modelo.

En términos técnicos, la función de la Devianza se define como:

$$d(Y_i; \mu_i) = 2\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)} d\zeta$$

Bajo la condición que $\mathrm{V}(\zeta)$ es estrictamente positiva y por tanto, $\mathrm{d}(\mathrm{Y_i}; \mu_\mathrm{i})$ es también estrictamente positiva y satisface la condición para ser una función de distancia. De esta forma, sumando dicha función de la Devianza a lo largo de todas las observaciones de la muestra de datos, da como resultado la medida de Devianza total denotada como:

$$D = \sum_{i=1}^{n} 2\,\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)}\, d\zeta$$

Siendo así, cuando el modelo se ajusta bien entonces se espera que el valor de la Devianza sea pequeño; ya que de lo contrario, indica que se trata de un modelo mal ajustado.

Ahora bien, el tamaño de $D$ se evalúa en relación a la distribución muestral. Por lo tanto, a manera de resumen, para miembros específicos de la Familia Exponencial se tendría:

La Devianza es más usada para comparar dos modelos que como medida de bondad de ajuste absoluta. Esto es, si se quiere contrastar dos modelos al añadir una nueva variable, la Devianza proporciona el nivel de mejora que proporciona al modelo *(Tabla 35)*.

| DISTRIBUCION | Devianza |
|---|---|
| ***Normal*** | $\sum_i w_i (y_i - \mu_i)^2$ |
| ***Poisson*** | $2\sum_i w_i \left[ y_i \, log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right]$ |
| ***Gamma*** | $2\sum_i w_i \left[ -log\left(\frac{y_i}{\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i} \right]$ |
| ***Binomial*** | $2\sum_i w_i m_i \left[ y_i \, log\left(\frac{y_i}{\mu_i}\right) + (1 - y_i) \, log\left(\frac{1-y_i}{1-\mu_i}\right) \right]$ |
| ***Inversa Gaussiana*** | $\sum_i \frac{w_i (y_i - \mu_i)^2}{\mu_i{}^2 y_i}$ |

**Tabla 35:** Funciones de Devianza
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

- Test de Wald

Cada vez que hay una relación dentro o entre los datos, se puede expresar un modelo estadístico con los parámetros a ser estimados a partir de una muestra. Pues

bien, el Test de Wald es una prueba estadística paramétrica que se utiliza para poner a prueba el verdadero valor del parámetro basado en la estimación de la muestra.

En este test, la estimación de Máxima Verosimilitud $\hat{\theta}$ de cierto parámetro determinado $\theta$, se compara con el valor propuesto $\theta_0$ bajo la suposición de que la diferencia entre ambos seguirá aproximadamente una distribución Normal.

Normalmente, el cuadrado de la diferencia se compara con una distribución de Chi-Cuadrada; siendo el estadístico de Wald a comparar:

$$\frac{\left(\hat{\theta} - \theta_0\right)^2}{var(\hat{\theta})}$$

Alternativamente, la diferencia también puede ser comparada con una distribución Normal; por lo que el estadístico de Wald quedaría:

$$\frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$$

donde $se(\hat{\theta})$ es el error estándar de la estimación de Máxima Verosimilitud.

El Test de Wald sobre varios parámetros a la vez se lleva a cabo utilizando una Matriz de Varianza. Así mismo, se puede utilizar en una gran variedad de diferentes modelos, incluyendo modelos que consideren variables tanto dicotómicas como variables continuas.

- Criterio de Información de Akaike *(AIC)*

El Criterio de Información de Akaike (*AIC*) más que ser una medida de bondad de ajuste de modelo estadístico, proporciona un método para la selección del modelo. Dado una muestra de datos, el *AIC* ayuda a "ordenar" los modelos candidatos; comparándolos entre ellos y por tanto poder elegir alguno.

El método de *AIC* no proporciona información sobre la calidad del modelo en sentido absoluto. Esto es, si todos los modelos candidatos encajan mal, *AIC* no dará ninguna alarma sobre dichos modelos. Es decir, sólo ofrece una estimación relativa de la información perdida bajo cierto modelo determinado.

En términos generales, su definición está dado por:

$$AIC = 2k - 2\ln(L)$$

donde $k$ es el número de parámetros en el modelo estadístico y $L$ es el máximo valor de la función de Máxima Verosimilitud para el modelo estimado.

Para un conjunto de modelos candidatos para la muestra de datos que se tiene, el mejor modelo que ajusta, será aquel con menor valor *AIC*. Por lo tanto, *AIC* no sólo ayuda en la bondad de ajuste del modelo, sino también sirve para evitar el *"sobreajuste"* del modelo (sobreentrenar un algoritmo de aprendizaje con datos que se conoce el resultado deseado).

- Leverage

Es un estadístico que se utiliza como medida de influencia. Esto es, provee información de cómo valores individuales pueden potencialmente afectar los resultados del modelo. Nos ayuda a identificar observaciones puntuales con excesiva influencia en el modelo, en otras palabras, datos atípicos en la muestra.

Su definición formal es compleja pero esencialmente, representa, para cada observación de la muestra, la distancia del valor conjunto de las covariables para dicha observación respecto al valor medio de dichas covariables en el conjunto de todas las observaciones de la muestra.

Estrictamente, su valor debe estar contenido entre 0 y 1. Un valor de Leverage cercano a 1 significa que si cierta observación de la muestra ha presentado una mínima variación de cambio, entonces el valor ajustado por el modelo se moverá teniendo casi la misma variación. Es decir, existe una alta influencia del dato observado de la muestra sobre su valor ajustado por el modelo.

### 5.2.10. Sobredispersión

A menudo, para datos de recuento, es decir en los que se tiene varios eventos grabados en las mismas unidades (por ejemplo, registrar 0 siniestros o a lo más 1 siniestro); la variación observada es mayor que ésta, ya que los eventos serán interdependientes. En términos técnicos, se dice que existe una determinada relación entre las medias y varianzas condicionadas. Sin embargo, habitualmente las varianzas condicionadas son superiores a la media, lo que se conoce como *"sobredispersion"*.

Esto tiene como consecuencia una infraestimación de los errores estándar de los coeficientes del modelo, y sus causas pueden ser, entre otras, la presencia de heterogeneidad no observada, o bien, el incumplimiento del supuesto de independencia de los sucesos. En otras palabras, se dice que los sujetos de un grupo pueden no ser homogéneos; esto podría ser corregido con variables adicionales, por ejemplo características inherentes a cada individuo, que se podrían introducir en el modelo que ayuden a explicar las diferencias entre cada miembro. Sin embargo, este tipo de información no suele estar disponible o ser de fácil acceso para su incorporación en el modelo.

Para ello, surge la necesidad de recurrir al modelo de *Efectos Aleatorios*. Dicho modelo es una de las soluciones, más compleja pero a la vez más satisfactoria, es donde se asume que el parámetro media, el cual se supone desconocida entre las observaciones de la muestra, tiene una distribución aleatoria.

Cada miembro de la familia de dispersión exponencial tiene una distribución de composición correspondiente, conocido como conjugado, que produce de forma analítica cerrada la distribución compuesta. De tal forma que, para una distribución de la Familia Exponencial, la distribución conjugada del parámetro aleatorio sería:

$$p(\theta;\zeta,\gamma) = exp[\zeta\theta - \gamma b(\theta) + s(\zeta,\gamma)]$$

donde $s(\zeta,\gamma)$ es un término que no implica al parámetro $\theta$.

De esta forma, esta distribución conjugada es también un miembro de la Familia Exponencial. Por lo tanto, la distribución del compuesto resultante, para $n$ observaciones, sería:

$$f(y;\zeta,\gamma) = exp[s(\zeta,\gamma) + c(y) - s(\zeta + y\,,\gamma + n)]$$

la cual ya no sería un miembro de la Familia Exponencial.

### 5.2.11. Residuos

En los modelos lineales clásicos, se asume que tanto la variable respuesta como los errores del modelo se distribuyen de forma Normal. No obstante, en los GLM, no todos los datos van a seguir una distribución Normal, ni tampoco van a presentar una varianza constante; por el contrario, muchos de ellos presentan estructura No Normal, por lo que los valores de la estimación del modelo van a seguir la misma distribución que los datos iniciales.

Para detectar la normalidad (o no) de nuestros datos, es conveniente conocer el tipo y naturaleza de nuestra variable respuesta y analizar los residuos del modelo una vez ajustado. Pues bien, los residuos se definen como la diferencia entre los valores observados y ajustados.

Los residuos se pueden utilizar para comprobar la adecuación de ajuste de un modelo, con respecto a la elección de la función de la varianza, la función de enlace y los términos en el predictor lineal. También pueden indicar la existencia de valores atípicos que requieren de un mayor estudio y manejo de éstos dentro de la muestra de datos.

Algunos de las distintas tipos de residuos que se utilizan son:

- Residuos de Pearson

Los Residuos de Pearson no se emplean tanto como una prueba de bondad de ajuste, sino más bien como una media de la variación residual. Éstos se definen como:

$$r_p = \frac{y - \mu}{\sqrt{V(\mu)}}$$

siendo el residuo "puro" con respecto a la desviación estándar estimada de $Y$

- Devianza Residual

Retomando el concepto de la medida de Devianza total denotada como la suma de la función de Devianza sobre todas las observaciones de los datos:

$$D = \sum_{i=1}^{n} d_i$$

$$D = \sum_{i=1}^{n} 2\,\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)}\, d\zeta$$

siendo $d_i$cada función de Devianza, se llega nuevamente a la medida de Devianza Total. A partir, de aquí se define a la Devianza Residual como:

$$r_t^D = sign(Y_i - \mu_i)\sqrt{2w_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)}\, d\zeta}$$

Que se traduce en la raíz cuadrada de la contribución de cada una de las observaciones al total de la Devianza, quien mide la distancia entre la observación y su valor ajustado por el modelo; multiplicado por 1 o -1 dependiendo de sí el valor observado es mayor o menor que el valor ajustado.

La Devianza Residual cuenta con algunas propiedades de gran utilidad:

- La distribución de los residuos resultantes estarán mucho más cerca de una distribución Normal que los residuos propios del modelo (definidos simplemente como la diferencia entre las observaciones reales y los valores esperados predichos por el GLM), ya que la Devianza calculada corrige el sesgo de las distribuciones

- Si no se cumple la hipótesis de normalidad de la distribución de los residuos, existe una cierta desviación de la Normal

- En distribuciones continuas, es posible que la distribución de los residuos tienda a distribuirse a una Normal. Esta es una de sus propiedades más útiles, siempre y cuando la distribución de la respuesta se ha especificado correctamente.

### 5.3. Los GLM en la Práctica

Una vez se ha presentado como se deben ser formulados los Modelos Lineales Generalizados, resumiendo sus principales características, componentes y estructura, se procede a comentar ciertas cuestiones y una serie de pasos a seguir en el momento de iniciar la construcción de un modelo sobre cierto conjunto de datos.

En esta fase, es importante tener en cuenta que no existe un único modelo válido que se pueda ajustar a la muestra de datos analizados. Es decir, la mayoría de las veces, existe más de un modelo posible; es por ello, que el tema más complicado es saber y comprobar cuál ajusta mejor y por lo tanto, es el más adecuado de todos ellos.

Para ello, se pueden identificar cuatro fases que permiten estructurar la construcción de un GLM, que serían:

- Análisis Preliminar: El cual considera la etapa de preparación de los datos, así como un análisis exploratorio de las variables que serán consideradas dentro del modelo

- Iteración del Modelo: En esta fase, se recurre a la selección adecuada de factores que mejor se ajustan a los datos; y por tanto, hacer uso recurrente del diagnóstico de las hipótesis y parámetros del modelo

- Depuración del Modelo: Implica refinar el modelo buscando la máxima simplificación posible, encontrando la interacción entre las variables y haciendo uso de la suavización de los resultados

- Interpretación de Resultados: Finalmente, lograr traducir los resultados que se obtienen para obtener la mejor explicación del modelo e información que se obtiene del GLM planteado

### 5.3.1. Análisis Preliminar

Una de las primeras premisas que se debe tener en cuenta es la necesidad de datos. Un GLM es un modelo matemático intensivo en datos, el cual necesita un gran volumen de datos para poder obtener resultados coherentes y fiables. Esto en cuanto a número de observaciones, así como en la cantidad de variables que describan al evento analizado. En muchas ocasiones, para llegar a tener estos volúmenes de información, se recurre a lograr tener muestras de 2 o más años de exposición; esto es la misma observación analizada en distintos tiempos.

Ahora bien, antes de iniciar la modelación de los datos, es útil realizar cierto tipo de análisis preliminares de la información que se tiene. Esto incluye la identificación de valores "nulos" o vacíos, valores negativos donde valores "no lógicos", que teniendo cierto conocimiento de la naturaleza de la variable, no sea del todo adecuado ciertos valores que toma dicha variable. Para ello, los análisis sugeridos serían:

- Análisis de Distribuciones

Un análisis clave que ayuda a la identificación de características inusuales o contradicciones en los datos se puede obtener mediante la distribución de los datos (frecuencias, número de pólizas, importe de siniestros). Todo ello, con el fin de encontrar ciertas irregularidades o valores atípicos que deban ser analizados por aparte, o bien requieran cierto tratamiento previo a su modelación.

- Análisis Univariantes y Bivariantes

A pesar de que un GLM es un método multivariante, no deja de beneficiarse del análisis univariante y bivariante previo a la modelización; es decir, analizando variable por variable de manera independiente al resto, y posteriormente iniciar un análisis tomando de dos en dos variables.

Primeramente, empezar con el análisis de la distribución de los datos por cada una de las variables, proporcionaría información sobre el volumen de información que se tiene por variable. Con ello, tomar decisiones sobre si incluir o no la variable dentro

del modelo (por ejemplo, si más del 90% de la muestra recae en cierto valor de alguna de las variables propuestas, indicaría que no es una variable susceptible de modelarla).

Ahora bien, una vez obtenida la información sobre la distribución de los distintos niveles que toma cada una de las variables; es necesario saber si dichos niveles se combinan con otros; si existen similitudes, correlaciones o interdependencias entre ellas (ya que de existir algo de ello, sería muy complicado obtener una estimación de máxima verosimilitud).

- Categorización de Factores

Previo a la construcción del modelo, es necesario considerar como se debe categorizar las variables explicativas y cómo van a ser tratadas: de forma continua o de forma categórica o discreta. La mayoría de las veces es mejor ésta última opción; siempre y cuando se tenga suficiente volumen de datos por categoría, ya que el modelo los traducirá en factores categóricos estimados para ser empleados en forma de polinomios.

Por otro lado, es importante tener en cuenta la forma en como dichos factores son categorizados, ya que por lo general, se recurre a la combinación de niveles y factores. Por ejemplo, la categorización de la variable *"Edad"*, generalmente se busca tener masa suficiente en cada grupo de Edad, sin embargo si esto no es posible, se recurre a combinar esta variable con la variable "*Género"*; de tal forma que se acumule un mayor volumen de datos por categoría.

No obstante, la forma más apropiada de categorizar las variables viene dado por la propia estructura inicial de los datos; esto es, aquellos niveles de factores con poca exposición deberán ser agrupados en una misma categoría, y por el contrario aquellos con suficiente volumen entonces podrán ser considerados por separado.

- Análisis de Correlaciones

Una vez se tiene la categorización de los factores, es de gran utilidad conocer el grado de correlación que existe entre estos factores. Existen varios estadísticos de correlación para variables categóricas[37]; sin embargo, uno de los más utilizados dentro de la metodología de los GLMs, es el estadístico *V de Cramer*.

Primeramente, se define el Coeficiente *Chi-Cuadrado* como:

$$\chi^2 = \sum_{i,j} \frac{\left(n_{uj} - \frac{n_u\, n_j}{n}\right)^2}{\frac{n_u\, n_j}{n}}$$

siendo:

$n_{uj} =$ Número de observaciones para el *i-ésimo* Factor y para el *j-ésimo* Factor

Ahora bien, a partir de éste, el Coeficiente *V de Cramer* es una medida estadística de correlación entre dos factores categóricos, el cual se define como:

$$V = \sqrt{\frac{\chi^2}{n * (\min((a-1),(b-1))}}$$

donde:

$a =$ Número de valores del Primer Factor

$b =$ Número de valores del Segundo Factor

$n =$ Número de observaciones de la muestra total

Por último, El Coeficiente *V de Cramer* se puede interpretar considerando que su resultado se encuentra dentro del rango de valores del [0 – 1]:

- *V de Cramer = 0* → No hay relación entre los Factores
- *V de Cramer = 1* → Existe una relación perfecta entre los Factores 1

[37] Coeficiente de correlación de Pearson, el Test de Chi-Cuadrado, Coeficiente de Contingencia o el Coeficiente Phi o también llamado Coeficiente de Correlación de Mathews

- *V de Cramer = 0.6* → Hay una correlación relativamente intensa entre ambos Factores

En el modelo GLM, es importante comprender las correlaciones dentro de las variables de la muestra en el momento de interpretar los resultados; ya que puede ser de gran utilidad en la identificación de factores más afectados por la adición o eliminación de factores adicionales en el modelo. Además de ello, el análisis de la información de las correlaciones existentes, ayuda a detectar la existencia de colinealidad entre variables, que pudiese afectar al modelo.

### 5.3.2.- Iteración del Modelo

Una vez se tienen los datos observados analizados y se han estimado los parámetros que cumplen las hipótesis del modelo; se debe encontrar los parámetros que mejor ajustan a los datos, es decir la estimación más adecuada del modelo. Sin embargo, esto no es algo trivial y que se obtenga de forma directa e inmediata del modelo, ya que esta labor se deberá realizar a partir de la iteración del modelo una serie de veces. Es en esta fase donde, a partir de una selección adecuada de los factores, se procede al diagnóstico y validación de las hipótesis estadísticas asumidas, lo cual se realiza mediante los siguientes pasos propuestos:

- Selección de Factores

Para la elección de las variables explicativas que deberán ser consideradas en el modelo, se debe decidir de acuerdo al efecto que éstas tengan dentro del modelo. Para esta distinción, se cuentan con algunos métodos que se resumen a continuación:

❖ *Errores Estándar*

Aunque teóricamente, este tipo de prueba se puede hacer sobre las distintas estimaciones del parámetro, en la práctica suele ser más útil considerar, para cada factor, los parámetros junto con su error estándar asociado al nivel de base. Esto gráficamente se podría analizar de la siguiente forma (*Gráfica 1*):

- La línea verde indica los parámetros estimados ajustados para cierta variable

- La línea naranja muestra los resultados obtenidos del análisis univariante para esta variable. La diferencia con respecto a la línea verde se explica por las correlaciones entre variables: si ambas líneas siguen trayectorias muy dispares, significaría que la variable en cuestión está muy correlacionada con el resto de variables, y por lo tanto, su inclusión o eliminación afecta al resto; si por el contrario, las trayectoria son similares, entonces su influencia es nula.

- Finalmente, las líneas azules indican los errores estándar a ambos lados del parámetro estimado con un intervalo al 95% de confianza. Ahora bien, si estas líneas están muy juntas, indicaría que el parámetro es muy significativo; por el contrario, se encuentran muy separadas, indicaría gran incertidumbre en el parámetro estimado (debido a poco volumen o correlación de otro factores que explican mejor el riesgo)

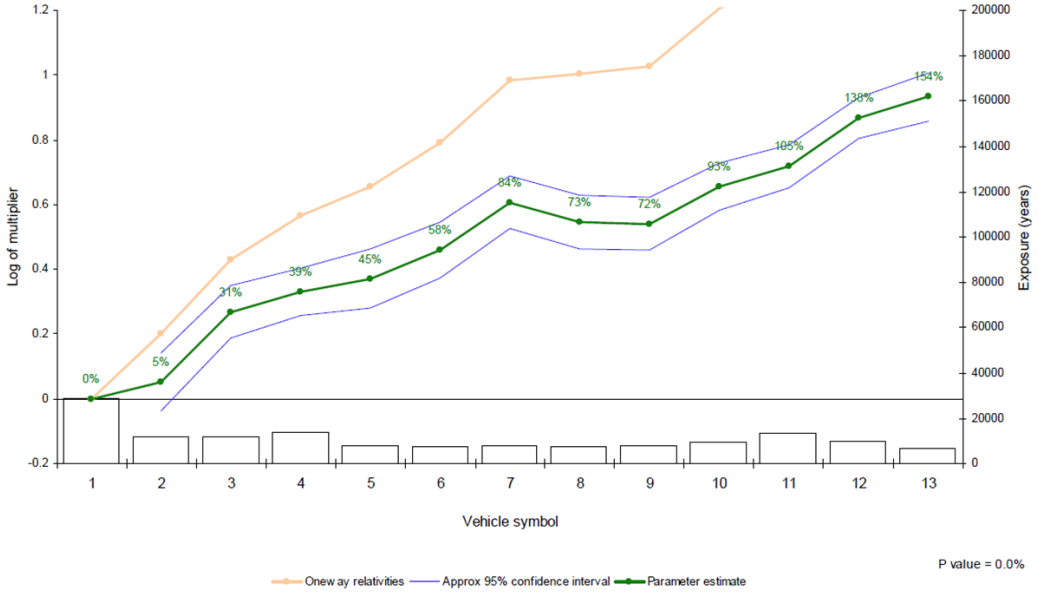


**Figura 39:** Ejemplo de *"Estimación de un Factor"* junto con sus *"Errores Estándar"*
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Aunque los errores estándar en el gráfico solo indican la certeza de la estimación del parámetro asociado a su nivel de base, es de gran ayuda para dar información sobre la significancia de las variables. Por ejemplo, en la *Figura 39* se puede deducir que el factor es significante. Por el contrario en la *Figura 40,* donde se analiza la misma variable pero dentro de un modelo diferente, se podría concluir que la variable no es significante.

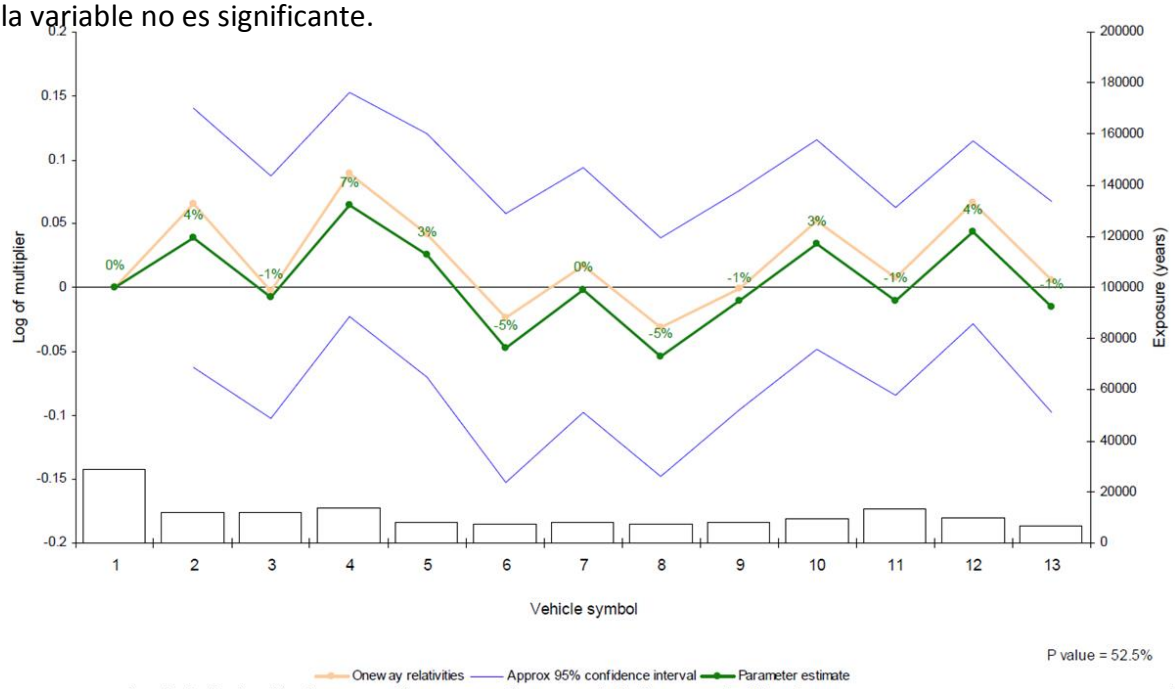


**Figura 40:** Ejemplo de *"Estimación de un Factor"* junto con sus *"Errores Estándar"* con poca significancia
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

- *Pruebas de Devianza*

Retomando, el término de Devianza mide la cantidad de varianza explicada por el modelo. Esta prueba arroja una idea de la variabilidad de los datos. A su vez, esta medida se ayuda de los *p-valores* de cada factor; quienes determinan la significancia teórica de los factores individuales. Si el *p-valor* es menor que o igual a 5%, entonces se puede considerar que la variable es significativa.

- *Interacción con el tiempo*

Por otro lado, podría ser de gran ayuda, utilizar técnicas más prácticas, como el revisar la consistencia de cierta variable a través del tiempo. Esto es, dividir la muestra de datos por año de exposición; y poder observar el comportamiento de cierta variable por cada uno de estos años. A partir de este análisis, poder elegir aquellos factores o variables que son consistentes a través de los años; y por tanto, se podría deducir que son buenos predictores de la futura experiencia del riesgo o evento analizado.

- *Sentido Común*

Adicionalmente a las tradicionales pruebas estadísticas de análisis exploratorio de variables; una técnica esencial en todo tipo de análisis es la intuición o sentido común. Esto es, comprobar de forma lógica si el efecto observado de un factor es similar al efecto esperado de acuerdo a modelos similares (tendencias lógicas, donde deben decrecer, etc.)

- Iteración del Modelo

Generalmente, no es posible determinar directamente desde un solo modelo GLM, que conjunto de variables son significativas; ya que con la inclusión o eliminación de cierto factor, podría cambiar los efectos y significancias de los factores correlacionados a éste dentro del modelo. Por lo tanto, es necesario llevar a cabo una serie de iteraciones del modelo para lograr determinar el conjunto de variables óptimas.

A menudo, la iteración del modelo inicia con un GLM que incluye todas las variables explicativas. A partir de aquí, se van excluyendo los factores que resulten insignificantes, uno a la vez, re-ajustando el modelo una y otra vez. Cuando se identifica un factor poco significante, es bastante útil recurrir a su análisis univariante, para determinar el modo en que su eliminación influirá en el resto del modelo.

Cuando existen una gran cantidad de variables posibles que considerar dentro de un modelo, puede ser complicado iniciar el proceso de iteración considerando todos y cada una de las variables explicativas. En estos casos, es recomendable seleccionar un conjunto de factores que se consideren importantes; y de manera

inversa, incluir uno a uno de los factores excluidos al modelo y medir el nivel de significancia que se gana con dicha inclusión.

En la medida de lo posible, es mejor poder iterar el modelo de forma manual, analizando en cada paso:

- ❖ La significancia de cada factor en el modelo, eliminando cada vez el menos significativo, considerando un cierto rango o límite mínimo de aceptación determinado por el usuario

- ❖ La significancia de cada factor no incluido en el modelo, comparándolo con un nuevo modelo que sí contenga el factor potencial a incluir

- ❖ Repetir estos 2 pasos hasta que todos los factores resulten significativos y todos los factores no incluidos en el modelo sean los de poca significancia

- ▪ Validación del Modelo

Además de tener en cuenta la significancia de los factores modelados, existen otras pruebas de diagnóstico del modelo, las cuales permiten la adecuación de otros supuestos del modelo que deben evaluarse. Para ello, se cuentan con algunos métodos que se resumen a continuación:

- ❖ *Residuos*

Se pueden obtener varios tipos de residuos para analizar cómo los valores esperados varían de los observados. Se suelen utilizar los residuos estandarizados y conviene analizarlos a través de los siguientes gráficos:

- Histograma de los Residuos

- Gráfico de los residuos vs los valores estimados para diagnosticar falta de linealidad y valores atípicos (Duncan Anderson, et. al. 2007)

- Gráfico probabilístico de normalidad (“*q-q plot”*), para contrastar la normalidad de la distribución de los residuos.

Después de revisar los resultados, quizá sea necesario replantear el modelo utilizando una estructura de errores más adecuada, con otra función vínculo o bien eliminando datos atípicos que puedan distorsionar el análisis.

A modo de ejemplo, la *Figura 41* muestra los resultados de un modelo GLM, donde se puede observar que yendo de izquierda a derecha sobre la gráfica, la media general y la variabilidad de la devianza residual, se observan razonablemente constante, sugiriendo que la función varianza es apropiada.

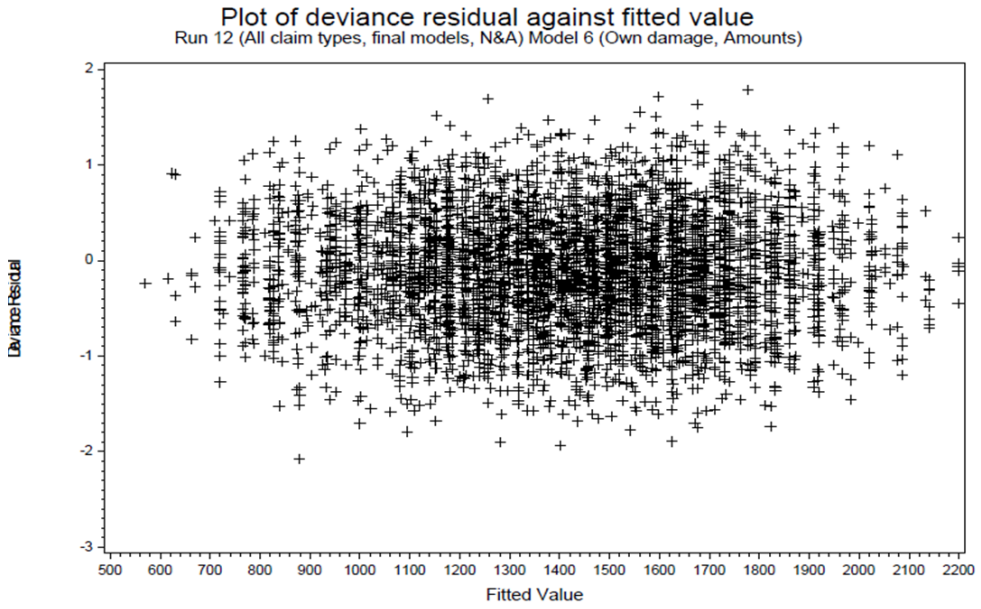


**Figura 41:** Gráfico de Residuos con un comportamiento contante
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Por el contrario, en la *Figura 42* de ejemplo, se observa que la variabilidad crece en la medida en que el valor ajustado lo hace, indicando que se ha seleccionado una función Error inapropiada; y por lo tanto, la varianza de las observaciones incrementa con el valor ajustado más de lo que ha sido asumido.

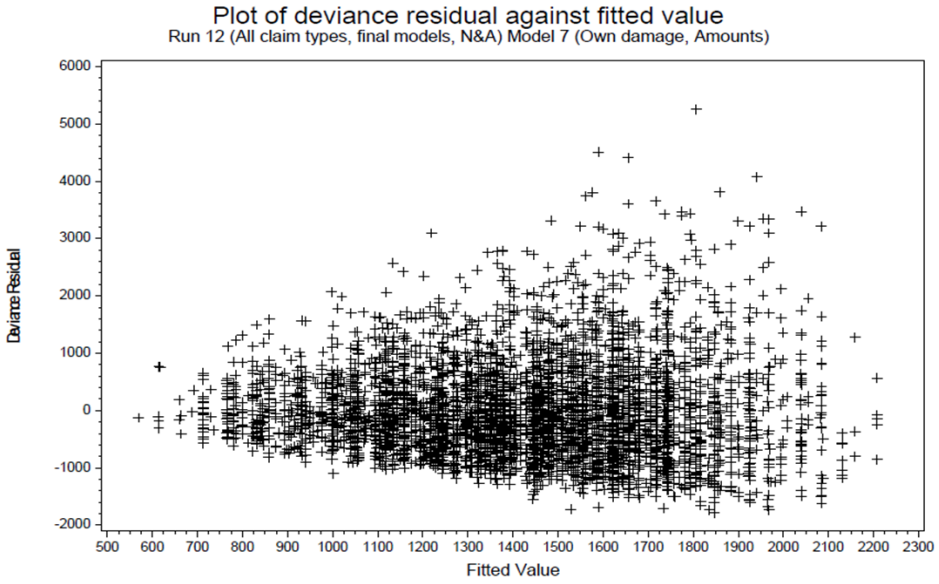


**Figura 42:** Gráfico de Residuos con un comportamiento irregular
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

❖ *Leverage*

Esta prueba del diagnóstico ayuda a identificar las observaciones que tienen influencia excesiva en el modelo; es decir, es útil en la identificación de observaciones puntuales con bastante influencia en el modelo, esto es, valores atípicos.

Nuevamente, un ejemplo en donde se pueden detectar, es mediante un gráfico del *Leverage (o Apalancamiento*) vs el valor ajustado. Así en la *Figura 43* se puede observar claramente algunas de las observaciones con un nivel de *Apalancamiento* muy por encima del comportamiento del resto, por lo que es evidente que estas observaciones están teniendo mayor influencia sobre los resultados del ajuste del modelo.

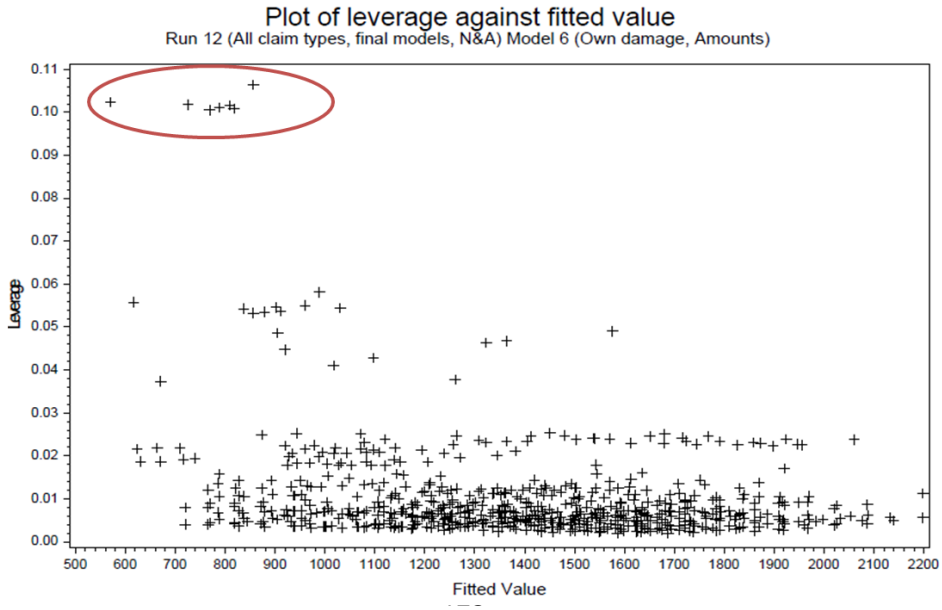


**Figura 43:** Gráfico de Leverage identificando valores atípicos
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Después de observar este gráfico, se podría plantear si es o no apropiado mantener estos datos atípicos dentro del modelo.

❖ *Transformación Box-Cox*

Comprueba la idoneidad de la Función Enlace seleccionada. Se define como la función link en términos de un parámetro escalar $\lambda$, siendo:

$$g(x) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases}$$

De aquí se desprende que los valores de $\lambda$ cercanos a *0,* sugieren que una estructura multiplicativa con una función de enlace logarítmica sería lo más apropiado para la muestra analizada. En la *Figura 44* se muestra un gráfico ejemplo de los resultados que podrían ser obtenidos para un análisis de Frecuencias de Siniestros, donde el $\lambda$ óptimo se encuentra cercano al cero. Por el contario, los valores de $\lambda$ cercanos a *1* sugieren que una estructura aditiva sería mejor.

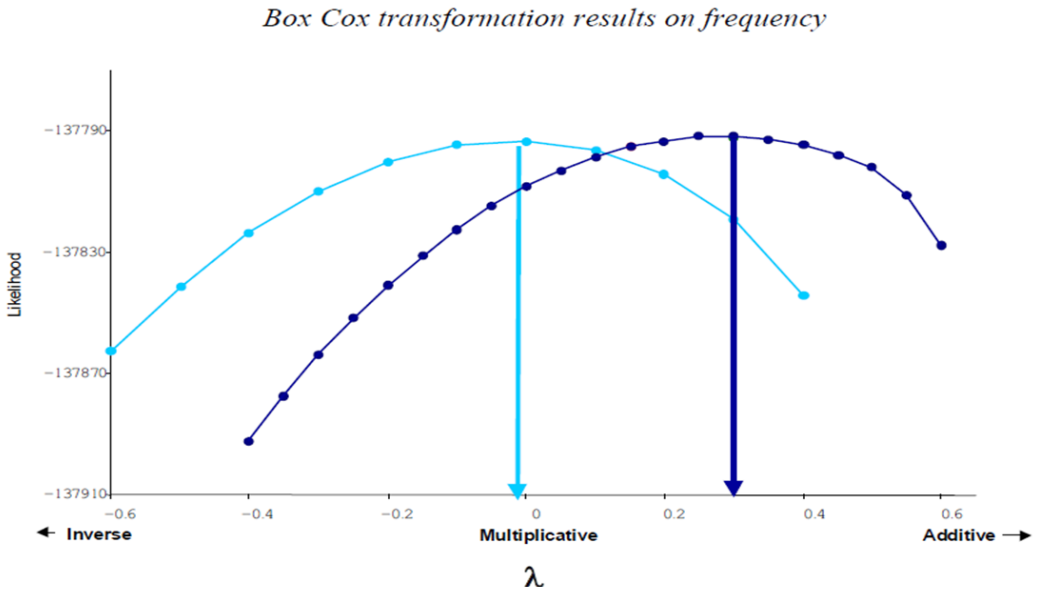


**Figura 44:** Gráfico de Transformación de Box-Cox para los resultados de un Modelo de frecuencias
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Finalmente, los valores cercanos a *-1,* indicaría que lo más adecuado es una función de enlace inversa; esto el caso mostrado en la *Figura 45,* el cual es un ejemplo donde también se encuentra cercano al cero; sin embargo, con cierta tendencia hacia el valor *-1*.

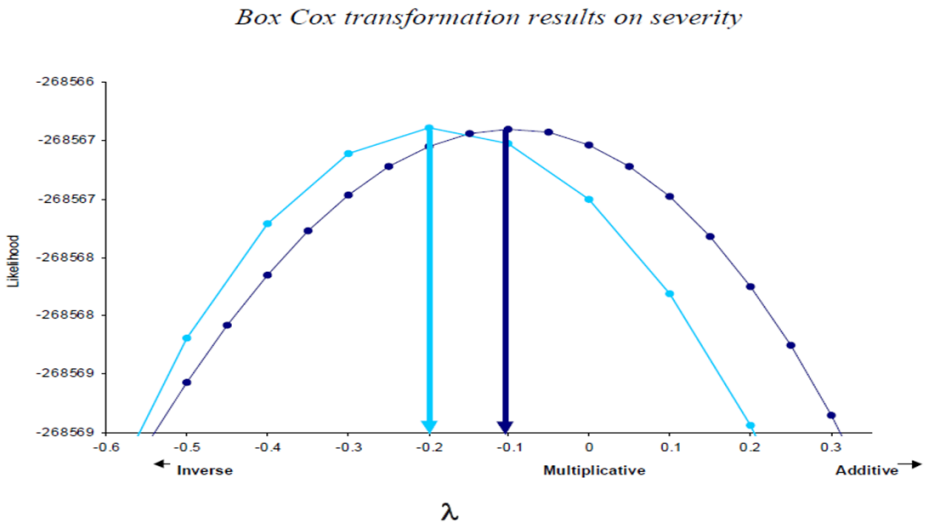


**Figura 45:** Gráfico de Transformación de Box-Cox para los resultados de un Modelo de Severidad
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

### 5.3.3. Depuración del Modelo

Los GLM se rigen por el principio de parsimonia, que intenta la máxima simplificación posible del modelo; es decir, que el modelo utilice el mínimo de factores para explicar el evento. Lo cual significa que no deberían existir parámetros o niveles de factor redundantes. Es por ello que surge la necesidad de tener una depuración o simplificación del modelo propuesto. Esto se realiza mediante 2 procesos identificados:

- *Interacciones*

Estas se dan cuando el efecto de un factor varía con los niveles de otro factor. Las interacciones hacen referencia al efecto que los factores tienen en el riesgo, y no a la correlación entre ellos. Son incluidas en el modelo mediante variables compuestas por dos o más variables (por ejemplo, en lugar de considerar *edad* y *sexo* por separado, se podría considerar como una única variable *edad-sexo*)

Las interacciones se deben incluir siempre y cuando exista una justificación estadística de su inclusión. En términos generales, esto se puede responder mediante la evaluación de la significancia de las interacciones:

- Evaluando los Errores Estándar del parámetro estimado del término marginal
- Mediante el *p-valor* del término marginal
- Con la consistencia con la interacción con el tiempo

- *Suavizado*

Una vez que se ha concluido con la iteración de los modelos y han sido incluidas las interacciones propuestas, se procede a la suavización de los parámetros en busca de una mejora en el poder predictivo del modelo. En esta fase, se incorpora ciertos conocimientos o juicio del experto en el evento modelado. Es decir, el experto deberá aportar cierto conocimiento buscando un comportamiento natural del evento estimado

Para el suavizado de un modelo no existe algún marco "científico" que aporta ciertas reglas de actuación, sin embargo, es de tener en cuenta que, el experto, debe preguntarse si se debe suavizar o bien replantearse ciertos parámetros o hipótesis del modelo en su estructura propia.

### 5.3.4. Interpretación de Resultados

Esta es la fase más complicada pero a su vez la más enriquecedora, ya que con base en ella, se puede entregar resultados para una correcta toma de decisiones, o bien, no lograr interpretar los resultados de una forma atractiva y que aporte valor.

Así también, esta fase recurre a los gráficos para la interpretación de los resultados de una manera más ágil. La mayoría de los gráficos, muestran cómo afecta un determinado factor o variable al Nivel base. Algunos estos tipos de gráficos, a manera de ejemplo, serían *(Figura 46)*:

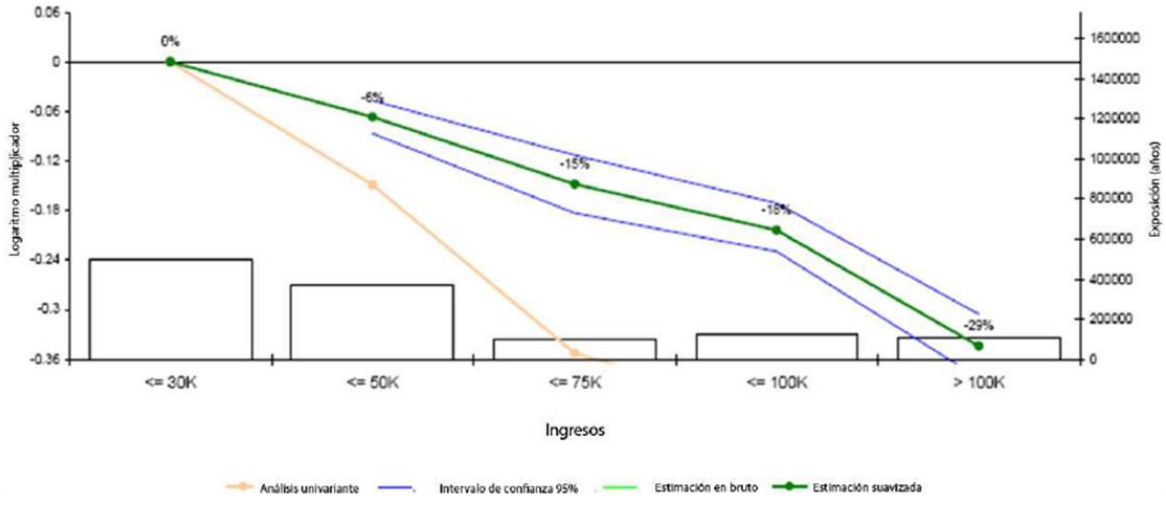


**Figura 46:** Gráfico de Impacto de la variable Suma Asegurada
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

En la *Figura 46* se observa cómo impacta la suma asegurada en el evento modelado; indicando que dicho evento desciende según aumenta el importe de suma asegurada (línea verde que da los resultados del GLM)

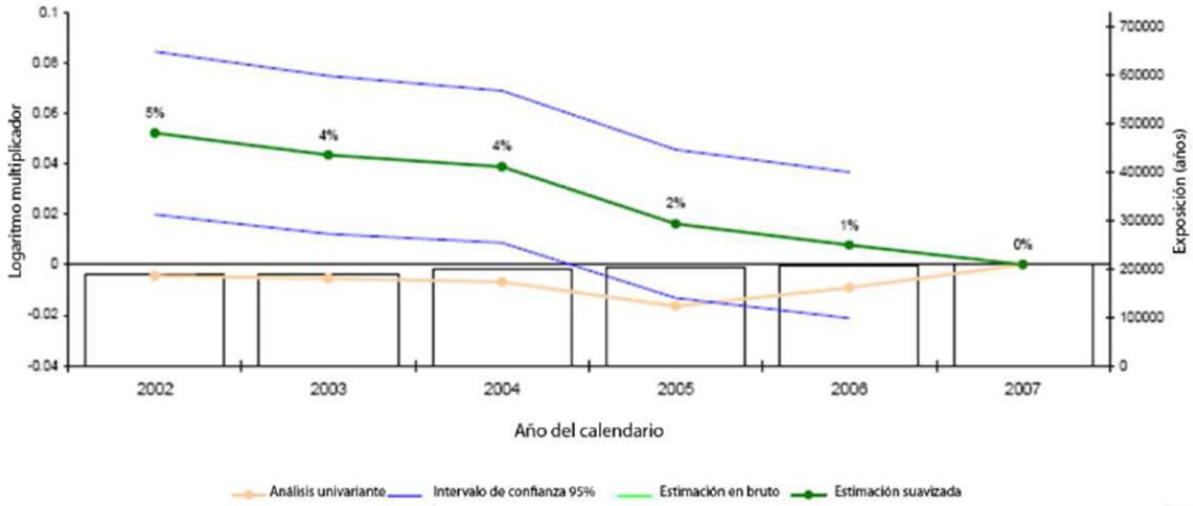


**Figura 47:** Gráfico de Impacto de la variable Año Calendario
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Otro ejemplo de interpretación, sería el siguiente gráfico donde se observa el impacto del año calendario sobre el evento modelado. En la *Figura 47* el Nivel Base es el año calendario 2007, donde se observa cierto descenso en la ocurrencia del evento con el transcurso de los años.

Así mismo, en ambos ejemplos, se puede observar un intervalo de confianza del 95% para la significancia estadística, donde visualmente, se puede concluir que entre más ancho sea dicha "franja", el factor tiende a ser menos significativo. Y finalmente, la cantidad de exposición o número de observaciones para cada categoría viene determinada por el gráfico de barras sobre el eje $x$.

### 5.3.5. Ventajas y Limitaciones

Primeramente, haciendo referencia a sus bondades, se remarca el hecho de un GLM ofrece una forma relativamente simple y robusta para analizar el efecto de distintos factores sobre un evento observado.

Por otro lado, es un modelo que permite analizar las interacciones entre los factores, esto es, se lograr estudiar cierto evento considerando varios factores que

influyen en éste a la vez; lo cual es útil en el entorno actual donde se sabe que más de un factor influye en los riesgos o eventos analizados con frecuencia. En términos generales, los modelos GLM son robustos, transparentes y de fácil comprensión e interpretación-.

Sin embargo, pese a su notable generalidad, los GLM no están exentos de limitaciones. Una de ellas deriva de su carácter de modelos "lineales". Esto es, las variables explicativas $x_i$ entran en el modelo a través del predictor lineal $\varphi$ que no es más que una combinación lineal de esas variables explicativas, esto es: $\varphi = \sum_i \beta_i \, x_i$

Pues bien, una forma de introducir efectos no lineales en el modelo es discretizando las variables $x_i$ ya que con frecuencia, dichas variables son continuas. Por ejemplo la edad del tomador de seguros, se suele discretizar para convertirlas en categóricas. La discretización o categorización de las variables explicativas permite tener en cuenta de forma sencilla, aunque no muy precisa, posibles efectos no lineales de las variables explicativas sobre las variables dependientes. Pero tiene inconvenientes como la pérdida de información; es decir, existe una cierta arbitrariedad a la hora de establecer los límites que determinan cada una de los rangos categóricos, o bien la existencia de rangos con pocas observaciones para las que resulta difícil obtener estimaciones precisas.

Otra de sus limitantes proviene de las restricciones derivadas de los datos. Esto es, un GLM es un modelo intensivo en datos lo cual requiere de una suficiencia y certeza en la información utilizada. Finalmente, un GLM ofrece una plataforma para modelizar más que una respuesta en sí misma, por lo que se recurre a una gran cantidad de juicios de valor para la interpretación de los resultados y análisis de tendencia a partir de estos modelos.

### 5.4. Aplicación Empírica

Llegado a este punto, se considera importante, hacer una recapitulación de los dos principales conceptos que están siendo tratados: que serían:

- Riesgo de Caída de Cartera: Se enmarca dentro de las exigencias del nuevo marco de Solvencia II, siendo uno de los riesgos por cuantificar para ser considerado dentro del requerimiento de capital exigido a cada una de las entidades aseguradoras. Para lograr la "mejor estimación" de dicha cuantificación, se requiere de modelos predictivos que proporcionen mecanismos de alarma, medición y gestión de riesgos oportuna

- Modelos Lineales Generalizados: Son uno de estos modelos predictivos capaz de, primeramente, identificar un conjunto de factores de riesgo que explican el comportamiento y los cambios en la variable respuesta (evento analizado), y que posteriormente, lo convierten en un resultado numérico en función de los distintos niveles de dichos factores

Pues bien, en esta sección se abordará la aplicación empírica de un modelo predictivo como lo son los GLM, para la identificación de variables o factores que indiquen el posible abandono de una póliza de Vida en una entidad aseguradora.

Así mismo, dicha aplicación, más que buscar la exactitud de los resultados numéricos, busca ser una metodología contraste para las técnicas de Inteligencia Artificial. Es decir, mediante un modelo paramétrico como lo son los GLM se pretende complementar y fortalecer las conclusiones obtenidas del modelo no paramétrico que proporciona la Inteligencia Artificial.

Es así como la aplicación práctica sobre la que se desarrollará un modelo GLM será, al igual que se hiciese con la aplicación empírica de Inteligencia Artificial, sobre una muestra de pólizas de Seguros de Vida Individual.

Para ello, se intenta en la medida de lo posible, seguir las fases que la bibliografía (Duncan, Anderson et. al. 2007) sugiere para el desarrollo de un Modelo Lineal Generalizado. Como un primer bloque se buscará cubrir con el análisis preliminar de la información mediante los análisis univariantes y bivariantes donde se

plantea la posible relación o dependencia de dos en dos variables. Y posteriormente, se procederá a la selección óptima y aplicación empírica de un GLM.

Antes de comenzar, cabe aclarar que se trata de un ejemplo puramente teórico, que no está centrado en los resultados; ya que por un lado, se trata de una técnica que requiere de bastante experiencia en su uso, y además no es el objetivo de este trabajo. Si no se trata de contrastar los resultados obtenidos en las metodologías propuestas de Inteligencia Artificial con los factores identificados bajo la técnica GLM. Con ello se busca fomentar su uso como una alternativa novedosa o bien como un complemento a los actuales métodos utilizados en el sector, para que de manera conjunta, ofrezcan una gestión oportuna de los riesgos que la nueva regulación propone.

### 5.4.1 Análisis Preliminares

Como ya se ha dicho en cuanto al tema de la selección de variables, se hablaba de que pueden existir varios factores que influyen en el Riesgo de Caída de Cartera. Sin embargo, para efectos del estudio robusto de dicho riesgo, dicha selección de factores se podría ver limitada por el tamaño de la muestra. Sin embargo, es congruente con el objetivo del análisis, por lo que para conseguir esa consistencia en los resultados, se han seleccionado las mismas variables que han sido consideradas en la aplicación de las metodologías de Inteligencia Artificial.

Ahora bien, sólo mencionar que, de acuerdo a la metodología presentada de los GLM, en esta aplicación, la variable TIPO PRESTACION será la variable respuesta del modelo por plantear. Retomando, esta variable toma dos valores de acuerdo al estatus de la póliza *(Tabla 28)*:

| COD | TIPO_PRESTACION | Nº de POLIZAS | % Peso |
|---|---|---|---|
| 0 | VIGOR | 16,568 | *83.74%* |
| 1 | ANULADA | 3,216 | *16.26%* |
| **TOTAL** | | **19,784** | ***100.00%*** |

**Tabla 28:** *Distribución de acuerdo a la variable TIPO DE PRESTACION*
**Fuente**: *Propia de los autores*

De esta forma, el resto serán las variables explicativas que entrarán en el modelo y serán quienes expliquen el comportamiento del riesgo de Caída de Cartera actuando como los factores del modelo.

Por último, es útil utilizar la *Transformación de Datos* que se hizo para las metodologías de Inteligencia Artificial; es decir, se ha trabajado con la serie de "códigos" que interpretan o clasifican el conjunto de variables. De esta forma, se sabe que se parte de la misma muestra de datos; y aprovecharse de esta ventaja que aporta los GLM, al ser modelos potentes a la hora de trabajar con variables categóricas.

#### 5.4.1.1. Análisis Univariante

Como ya se había mencionado, aunque un GLM se engloba dentro de las técnicas multivariadas de la Estadística Paramétrica; no es por demás recurrir al Análisis Univariante que proporciona la Estadística Descriptiva en su aplicación más simple y burda. Esto es un análisis descriptivo de los datos, retomando conceptos de Media, Moda, como se distribuye la muestra de datos por variable, etc. Todo ello, con el fin de conocer el volumen y tipo de información con la que se cuenta.

Sin embargo, esta fase ya ha sido resuelta en el tercer capítulo, presentando la descripción de cada una de las variables, contexto y muestra con la que se trabajaría; tanto para la metodología no paramétrica como para esta. Por lo que, se aprovechará ese trabajo previo, para cubrir con esta etapa del tratamiento de la información; del cual se obtiene el análisis variable a variable con el que se inicia el estudio empírico.

#### 5.4.1.2. Análisis Bivariante

Retomando las fases del modelo, un paso previo a la modelización, es la realización de un análisis preliminar de la información. Una vez obtenido el análisis variable por variable, ahora es necesario conocer si existen similitudes, correlaciones o dependencias entre éstas.

Por lo tanto, se procede al análisis bivariante de la información; con lo cual se inicia con la obtención de la matriz de correlaciones. En la siguiente tabla, se resumen en amarillo las correlaciones entre todas las variables; y en verde los p-valores que muestran su significación:

| Matriz de correlación | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SEXO | EDAD_ACTUARIAL | ANTIGUEDAD | TIPO_PRODUCTO | RED | FORMA_PAGO | EDO_CIVIL | HIJOS | VALOR_CLIENTE | ICE | NIV_INGRESOS | NIV_ESTUDIOS | TIPO_PRESTACION |
| SEXO | | 0.0048 | -0.0004 | -0.0807 | -0.0511 | 0.0577 | 0.0373 | 0.0095 | 0.0063 | 0.0302 | 0.0162 | 0.0663 | -0.0095 |
| EDAD_ACTUARIAL | 0.4999 | | 0.0796 | -0.2782 | -0.0211 | 0.3185 | 0.1745 | 0.1608 | -0.0661 | 0.1512 | 0.1101 | 0.1713 | -0.0692 |
| ANTIGUEDAD | 0.9586 | 0.0000 | | 0.1560 | 0.0173 | -0.0769 | 0.0115 | -0.0293 | -0.0853 | 0.0188 | 0.0387 | 0.0839 | -0.4217 |
| TIPO_PRODUCTO | 0.0000 | 0.0000 | 0.0000 | | 0.0568 | -0.8064 | -0.0509 | 0.0151 | 0.0234 | -0.0694 | -0.0687 | -0.1074 | 0.0373 |
| RED | 0.0000 | 0.0030 | 0.0148 | 0.0000 | | -0.0643 | -0.0001 | -0.0166 | 0.0256 | 0.0148 | 0.0101 | -0.0007 | -0.0007 |
| FORMA_PAGO | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | | 0.0552 | 0.0057 | -0.0130 | 0.0737 | 0.0639 | 0.1036 | -0.0793 |
| EDO_CIVIL | 0.0000 | 0.0000 | 0.1072 | 0.0000 | 0.9834 | 0.0000 | | 0.1589 | 0.0441 | 0.0302 | 0.0313 | 0.0326 | -0.0134 |
| HIJOS | 0.1807 | 0.0000 | 0.0000 | 0.0341 | 0.0199 | 0.4188 | 0.0000 | | -0.0040 | 0.0140 | -0.0736 | -0.1540 | 0.0044 |
| VALOR_CLIENTE | 0.3742 | 0.0000 | 0.0000 | 0.0010 | 0.0003 | 0.0676 | 0.0000 | 0.5767 | | -0.0557 | -0.0189 | -0.0152 | 0.0676 |
| ICE | 0.0000 | 0.0000 | 0.0082 | 0.0000 | 0.0375 | 0.0000 | 0.0000 | 0.0491 | 0.0000 | | 0.0401 | 0.0636 | -0.0247 |
| NIV_INGRESOS | 0.0230 | 0.0000 | 0.0000 | 0.0000 | 0.1566 | 0.0000 | 0.0000 | 0.0000 | 0.0078 | 0.0000 | | 0.6043 | -0.0322 |
| NIV_ESTUDIOS | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9220 | 0.0000 | 0.0000 | 0.0000 | 0.0329 | 0.0000 | 0.0000 | | -0.0473 |
| TIPO_PRESTACION | 0.1795 | 0.0000 | 0.0000 | 0.0000 | 0.9202 | 0.0000 | 0.0586 | 0.5342 | 0.0000 | 0.0005 | 0.0000 | 0.0000 | |

De un simple vistazo, se observa como las variables están muy correlacionadas unas con otras, por lo que se espera que existan ciertas interacciones entre ellas. Es aquí donde entra una de las ventajas de los Modelos Lineales Generalizados, ya que éstos son capaces de tener en cuenta las interacciones entre las variables, que por el contrario técnicas de regresión lineal no lo consideran.

A partir de aquí, se puede iniciar con un estudio de variables dos a dos. Sin embargo, debido al gran número de variables explicativas consideradas en el modelo, esto da lugar a un considerable conjunto de combinaciones. Por lo que, a manera de resumen, se expone parte del análisis bivariante. Esto es, se analizará la correlación de cada una de las variables explicativas con respecto a la variable respuesta TIPO PRESTACION:

- SEXO
  - *Coeficiente de Correlación = -0.0095*
  - *P-valor = 0.1795*

Por un lado, el Coeficiente de Correlación es muy cercano a 0, por lo que se podría deducir que no existe una relación lineal entre el SEXO y la variable respuesta TIPO PRESTACION. Por otro lado, tomando un nivel de confianza del 90% siendo un poco conservadores; significaría que si el p-valor es mayor que el nivel de significación establecido del 0,1 (equivalente a 100%-90%), entonces se supondría que no existe dependencia entre las variables. Con lo cual, cabría suponer que la variable SEXO tendrá poca influencia dentro del modelo.

- EDAD ACTUARIAL
  - *Coeficiente de Correlación = -0.0692*
  - *P-valor = $<2.2e^{-16}$*

Nuevamente el Coeficiente de Correlación es muy cercano a 0, por lo tanto se deduce que no existe una relación lineal entre las variables. Sin embargo, según el p-valor, se observa que la variable EDAD ACTUARIAL y TIPO PRESTACION son

dependientes entre sí y sugiere que debería ser considerada en el modelo por su posible influencia en los resultados.

- ANTIGUEDAD

  - *Coeficiente de Correlación = -0.4217*
  - *P-valor = $<2.2e^{-16}$*

En este caso, el Coeficiente de Correlación es altamente significativo, encontrándose entre -1 y 0, lo que cabría indicar que existe una correlación negativa de las variables. Y mediante el p-valor cercano a 0, se confirma que se trata de una variable que debe ser claramente considerada en el modelo.

- TIPO PRODUCTO

  - *Coeficiente de Correlación = 0.0373*
  - *P-valor = $<2.2e^{-16}$*

Para esta variable, el Coeficiente de Correlación también es muy cercano a 0 pero por el lado positivo; una vez, se supone la existencia de cierta relación lineal entre el TIPO PRODUCTO y el TIPO PRESTACION. Así mismo, se considera que ambas variables son dependientes, ya que el p-valor es cercano a 0.

- RED

  - *Coeficiente de Correlación = -0.0007*
  - *P-valor = 0.9202*

Aquí se tiene una clara evidencia de una variable que, en principio, por los resultados del Coeficiente de Correlación que es muy cercano a 0, se asumiría que no relación lineal con respecto a la variable independiente. Finalmente, esto se confirma con la segunda prueba, donde el p-valor es muy cercano a 1, lo cual rechazaría todo nivel de significancia y por tanto, debería estar fuera del modelo por completo.

- FORMA PAGO

- *Coeficiente de Correlación = -0.0793*
- *P-valor = <2.2e$^{-16}$*

El Coeficiente de Correlación es cercano a 0, por lo tanto se deduce que no existe una relación lineal entre las variables. Sin embargo, según el p-valor, se observa que las variables son dependientes entre sí y por tanto, ser considerada en el modelo.

- ESTADO CIVIL
  - *Coeficiente de Correlación = -0.0134*
  - *P-valor = 0.0586*

En este caso, el Coeficiente de Correlación es nuevamente cercano a 0 por el lado negativo, entonces se deduce que no existe una relación lineal entre el EDO CIVIL y la variable respuesta TIPO PRESTACION. Por otro lado, el p-valor es un poco alto cercano al 1, por lo que en principio cabría decidir sacar la variable del modelo.

- HIJOS
  - *Coeficiente de Correlación = 0.0044*
  - *P-valor = 0.5342*

Muy similar a la variable anterior, donde el Coeficiente de Correlación es cercano a 0, y entonces supone que no existe relación lineal entre las variables. Y con el resultado del p-valor, cabría pensar en no incluir la variable HIJOS ya que muestra cierta independencia con respecto a TIPO PRESTACION.

- VALOR CLIENTE

  - *Coeficiente de Correlación = 0.0676*

  - *P-valor = $<2.2e^{-16}$*

Haciendo el mismo análisis, el Coeficiente de Correlación es muy cercano a 1 en este caso; lo cual asignaría cierta correlación positiva entre el VALOR CLIENT y la variable respuesta TIPO PRESTACION. Y esto se confirma con el p-valor, el cual indicia la dependencia entre las variables. En este caso, esto suena bastante lógico, ya que esta variable es asignada por la compañía en función del nivel de fidelización y rentabilidad que tiene el cliente con la entidad.

- ICE

  - *Coeficiente de Correlación = -0.0247*

  - *P-valor = 0.0005*

Aquí se tiene que el Coeficiente de Correlación es cercano a 0, marcando la no existencia de relación lineal entre las variables. Algo similar, es indicado con el p-valor, ya que éste también es cercano al 0 y muestra la dependencia entre ambas variables.

- NIVEL INGRESOS

  - *Coeficiente de Correlación = -0.0322*

  - *P-valor = $<2.2e^{-16}$*

Similar al caso justo anterior, donde el Coeficiente de Correlación indica la no relación lineal entre el NIVEL INGRESOS y TIPO PRESTACION. Y nuevamente, esto se confirma con el p-valor muy cercano a 0, con lo cual, cabría suponer que la variable NIVEL INGRESOS es dependiente a la variable respuesta.

- NIVEL ESTUDIOS

  - *Coeficiente de Correlación = -0.0473*
  - *P-valor = $<2.2e^{-16}$*

Finalmente observamos algo parecido con la variable NIVEL ESTUDIOS, con un Coeficiente de Correlación y p-valor cercanos a 0; con lo que se deduce nuevamente cierta dependencia entre las variables similar al NIVEL INGRESOS; lo cual hace sentido, suponiendo que, en cierta forma, ambas variables miden el nivel socio-económico del asegurado.

### 5.4.2. Aplicación del Modelo

La finalidad del presente estudio es lograr seleccionar un conjunto de factores óptimos que logren explicar el perfil del tomador de una póliza de seguros susceptible a la anulación de su contrato. Pues bien, para ello, primeramente se recurrirá a un análisis factorial de las variables cualitativas que se tienen. Esto con el fin nuevamente de conocer las relaciones y covarianzas que existen entre las variables explicativas; y proporcione más información sobre la muestra de datos que se tiene. Y ya no toda la información obtenida de los análisis en conjunto que se han presentado; se procederá finalmente a la aplicación y búsqueda de un Modelo Lineal Generalizado óptimo; y finalmente proceder a la elección de un Modelo mediante algunas pruebas de diagnóstico realizadas.

#### 5.4.2.1. Análisis Factorial

La finalidad de este análisis es resumir un gran número de variables en un número más pequeño de factores ficticios, creados a partir de combinaciones de distintos niveles de variables. Esto es, el Análisis Factorial es una técnica estadística de

reducción de datos usada para explicar las correlaciones entre las variables observadas en términos de un número menor de variables no observadas llamadas factores.

Ahora bien, este análisis es propuesto ya que al retomar la matriz de correlaciones (anteriormente expuesta); se puede observar que efectivamente las variables se encuentran correlacionadas entre sí y por tanto es recomendable aplicar dicho análisis.

Una de los modelos englobados dentro de esta técnica es el Análisis de Componentes Principales (su término en inglés *PCA – Principal Components Analysis*). De manera general, los PCA buscan reducir la dimensionalidad de un conjunto de datos, hallando las causas de variabilidad de un conjunto de datos y ordenándolas por importancia. Este método se utiliza para la construcción de modelos predictivos como es el caso del presente estudio. Como opera los PCA es mediante la descomposición de la matriz de covarianza en autovalores, normalmente tras centrar los datos en la media de cada atributo.

Aplicando la metodología sobre la muestra de datos que se tiene, los componentes que se obtiene son *(Tabla 36):*

| | | AUTOVALORES | | | |
|---|---|---|---|---|---|
| | | Autovalores | Diferencia | Proporción de Varianza | Proporción Acumulada |
| COMP | 1 | **2.18425** | 0.590533 | 0.1820 | **0.1820** |
| COMP | 2 | **1.59372** | 0.332198 | 0.1328 | **0.3148** |
| COMP | 3 | **1.26152** | 0.157099 | 0.1051 | **0.4200** |
| COMP | 4 | **1.10442** | 0.0659918 | 0.0920 | **0.5120** |
| COMP | 5 | **1.03843** | 0.0642665 | 0.0865 | **0.5985** |
| COMP | 6 | 0.974165 | 0.0414066 | 0.0812 | **0.6797** |
| COMP | 7 | 0.932758 | 0.0380378 | 0.0777 | **0.7574** |
| COMP | 8 | 0.894721 | 0.0946515 | 0.0746 | 0.8320 |
| COMP | 9 | 0.800069 | 0.15235 | 0.0667 | 0.8987 |
| COMP | 10 | 0.647719 | 0.266986 | 0.0540 | 0.9526 |
| COMP | 11 | 0.380733 | 0.193249 | 0.0317 | 0.9844 |
| COMP | 12 | 0.187484 | 0 | 0.0156 | 1 |

**Tabla 36:** Autovalores (Análisis de Componentes Principales)
**Fuente**: *Propia de los autores*

En el momento de elegir cuantos factores se quieren conservar, se puede seguir varios criterios:

- CRITERIO DE KAISER: El cual dice que los autovalores han de ser mayor que 1, ya que son los que más varianza explican. Bajo este criterio, la muestra se resumiría en 5 componentes, que son los que se encuentran por encima de 1.

- Otro criterio es elegir un mínimo de varianza a explicar. Esto es, dependiendo del porcentaje objetivo que se busque explicar con el modelo. Por ejemplo, siendo conservadores, ya que se sabe que la muestra que se tiene es totalmente empírica, se puede plantear un objetivo del 70-75%; y por lo tanto se consideraría quedarse entre 7 y 8 factores.

Finalmente, este tipo de análisis se rigen por el principio de la parsimonia, que dice que entre más simple sea el modelo, mucho mejor. Por lo que siguiendo este principio, se prefiere perder un poco de información, a cambio de tener menos factores, que nos hagan más interpretables los resultados. Es así como se elige quedarse con 7 componentes que serían (*Tabla 37)*:

| Autovectores | | | | | | | |
|---|---|---|---|---|---|---|---|
| | FACT1 | FACT2 | FACT3 | FACT4 | FACT5 | FACT6 | FACT7 |
| SEXO | 0.1033 | 0.0030 | 0.0119 | 0.1042 | -0.6528 | 0.5915 | 0.3450 |
| EDAD_ACTUARIAL | 0.4026 | -0.0352 | 0.3862 | -0.0853 | 0.0950 | -0.0752 | 0.0550 |
| ANTIGUEDAD | -0.0454 | 0.2488 | 0.3295 | -0.4114 | -0.0545 | -0.1298 | 0.5893 |
| TIPO_PRODUCTO | -0.5502 | 0.2779 | 0.2443 | 0.0342 | -0.0594 | 0.0267 | -0.0931 |
| RED | -0.0662 | 0.0797 | 0.0543 | 0.0689 | 0.7232 | 0.3976 | 0.3821 |
| FORMA_PAGO | 0.5570 | -0.2736 | -0.1897 | -0.0667 | 0.0720 | -0.0634 | 0.1405 |
| EDO_CIVIL | 0.1377 | -0.0642 | 0.5396 | 0.3788 | -0.0201 | -0.0691 | 0.0834 |
| HIJOS | 0.0059 | -0.2613 | 0.5280 | 0.1980 | -0.0548 | -0.1347 | -0.1970 |
| VALOR_CLIENTE | -0.0452 | -0.0519 | -0.1233 | 0.6877 | 0.1022 | 0.1705 | 0.1257 |
| ICE | 0.1505 | 0.0338 | 0.2377 | -0.3072 | 0.1229 | 0.6336 | -0.5132 |
| NIV_INGRESOS | 0.2646 | 0.5891 | -0.0205 | 0.1895 | 0.0046 | -0.1008 | -0.1663 |
| NIV_ESTUDIOS | 0.3067 | 0.5960 | -0.0374 | 0.1281 | -0.0332 | -0.0359 | -0.0606 |

**Tabla 37:** Autovectores de los Componentes Principales
**Fuente**: *Propia de los autores*

Para su interpretación, se observan las variables que saturan a cada factor; es decir las que más pesos tienen dentro de cada uno de ellos. Siendo así los 7 componentes elegidos quedarían explicados como:

- C1: Edad Actuarial – Forma de Pago – Nivel Estudios
- C2: Nivel Ingresos – Nivel Estudios
- C3: Edad Actuarial – Antigüedad – Edo Civil – Hijos
- C4: Edo Civil – Valor Cliente
- C5: Red
- C6: Sexo – Red – ICE
- C7: Sexo – Antigüedad – Red

#### 5.4.2.2. Elección del Modelo GLM

Una vez se han analizado las covariables, se procede a introducirlas en el modelo, para posteriormente iniciar la fase de análisis de los resultados. Pues bien, los GLM basan gran parte de su técnica en *"prueba y error",* de tal forma que van probando una serie de distribuciones y funciones link hasta encontrar cuál de ellas ajusta mejor los datos que se tienen. Sin embargo, se tiene conocimiento del tipo de muestra de datos que se está trabajando; donde el evento estudiado cuenta con dos opciones: se anula o se renueva la póliza de seguros; con lo cual la distribución que mejor ajusta a la variable respuesta será una binomial. Así pues, también se sabe que la función de enlace canónica para una binomial es *logit*, la que mejor funciona en la mayoría de las ocasiones.

Teniendo en cuenta todo lo anterior, el resumen de resultados obtenidos del primer modelo serían (con ayuda de la aplicación estadística *R Console*) *(Figura 48)*:

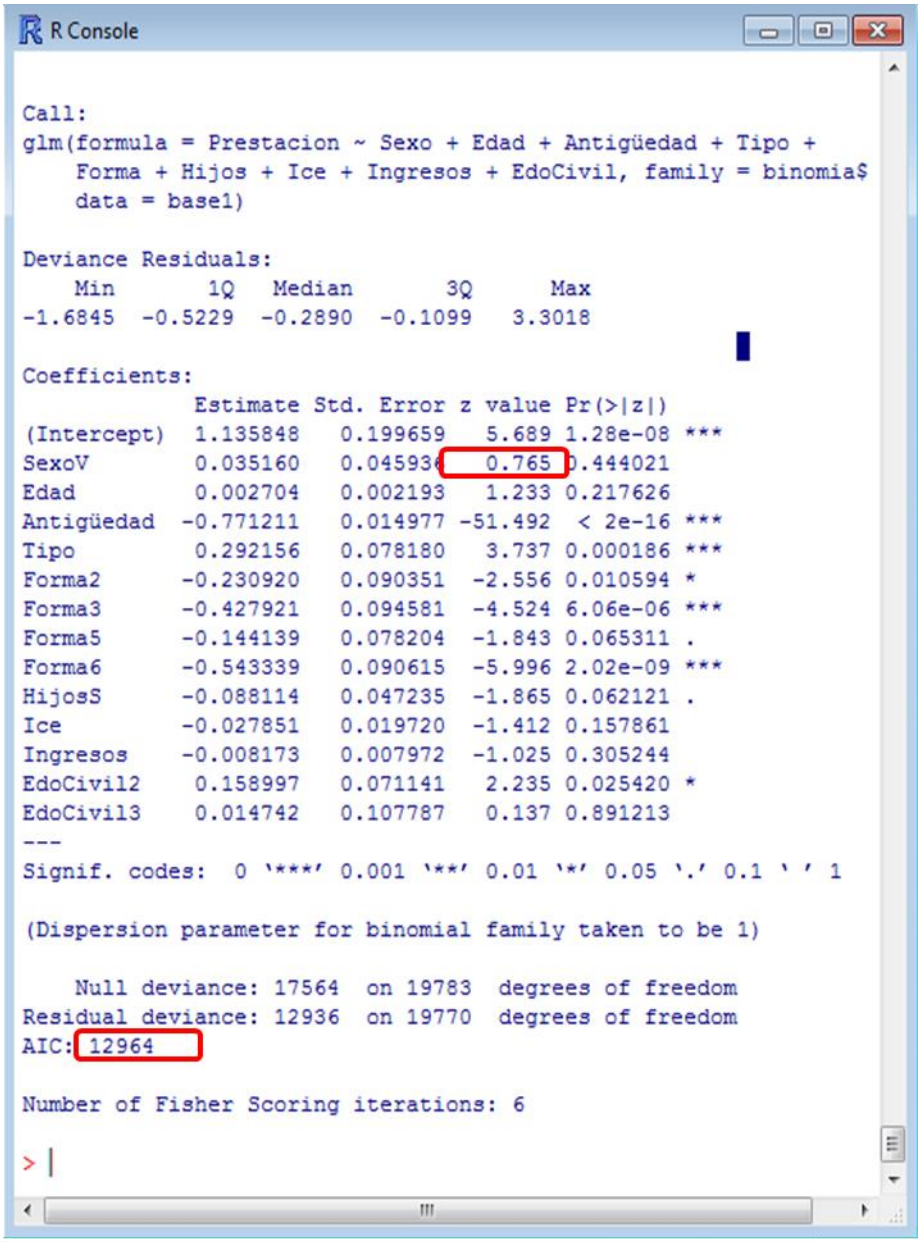


**Figura 48***: Salida de Resultados Modelo I – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Ahora se trata de ir quitando los factores menos significativos, hasta conseguir que todos o la mayoría de sus p-valor se encuentren por debajo de cierto intervalo de confianza. Inicialmente se puede considerar el 95% de confianza, aunque poco a poco seguramente se tenga que ser más conservador y bajar este intervalo de confianza.

Por otro lado, también se debe ir monitoreando como se mueve la medida de bondad de ajuste del Criterio de Información de Akaike (*AIC*); ya que nos

proporcionará un dato de referencia para la selección del modelo óptimo. Así se busca tener que dicho indicador deje de descender; con lo cual podría decir que se ha encontrado el modelo adecuado.

De esta forma, se decide primeramente quitar la variable SEXO, ya que tiene un *p-valor* equivalente al *0.765*; con lo cual los resultados del nuevo modelo sin esta variable quedarían *(Figura 49)*:

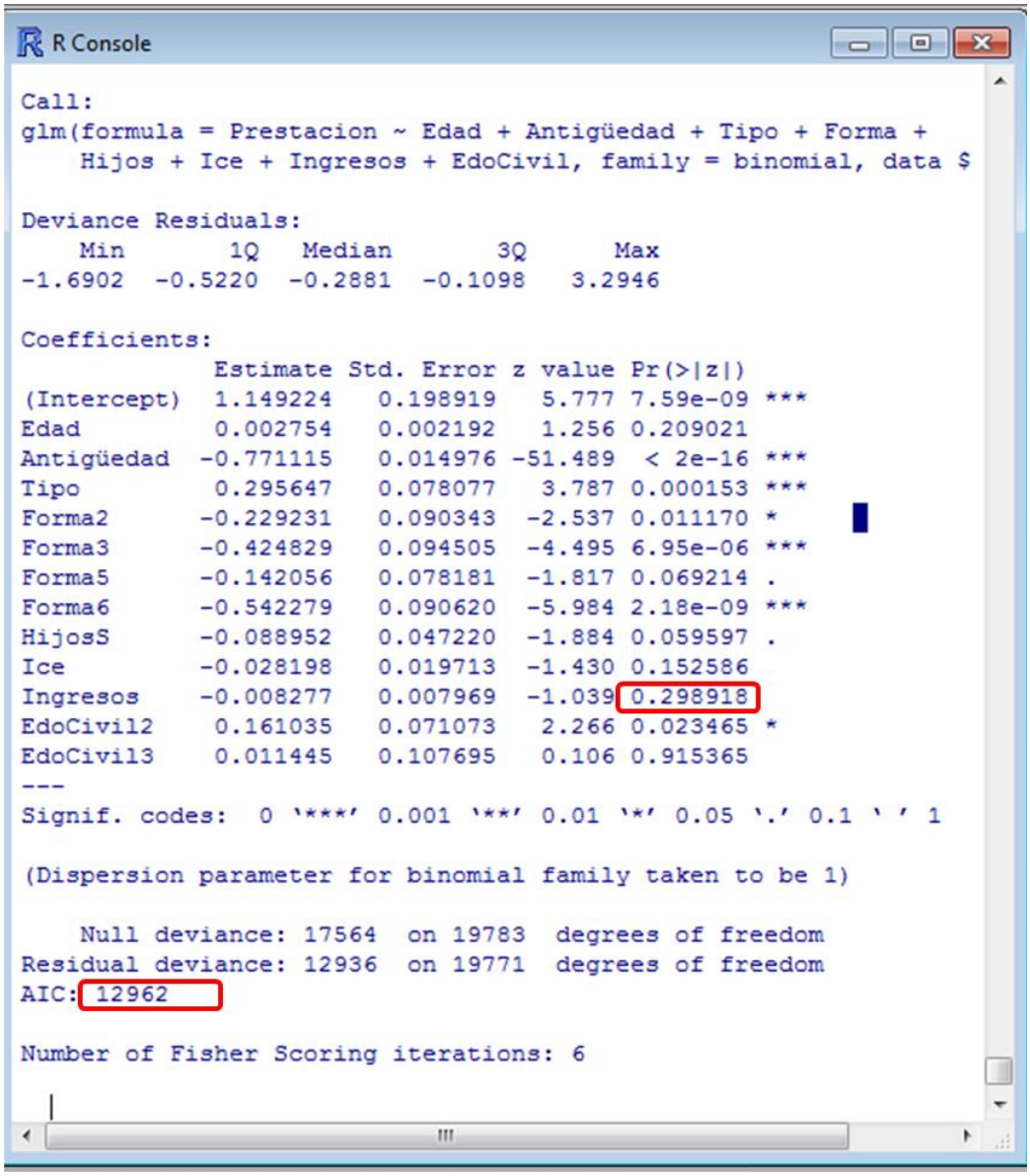


**Figura 49**: *Salida de Resultados Modelo II – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Se observa que el AIC se ha reducido y de la misma forma los p-valores. De esto se deduce que el modelo se ha mejorado eliminando la variable SEXO. Sin embargo, se observan p-valores muy por encima del objetivo inicial del 95%, por lo que se supone que se podría mejorar aún más el modelo con un nuevo intento.

En este caso, se decide sacar del modelo la variable INGRESOS, con lo que el resumen de resultados quedaría *(Figura 50)*:

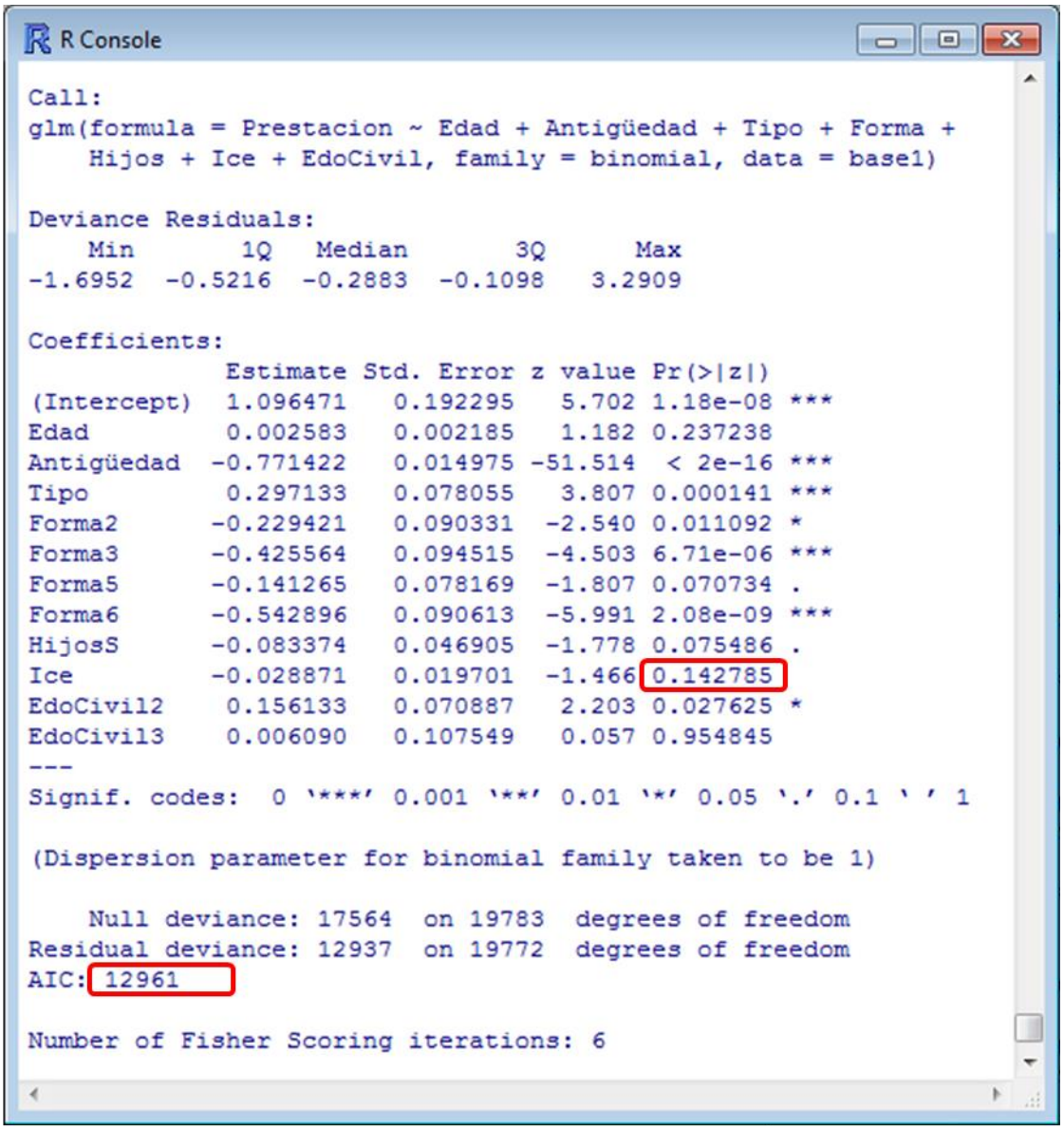


**Figura 50**: *Salida de Resultados Modelo III – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Nuevamente, se ha logrado reducir el AIC, por lo que se propone eliminar del modelo la variable ICE, quien es otro de los candidatos con un p-valor significativo.

Con esta nueva elección se tienen los siguientes resultados *(Figura 51)*:

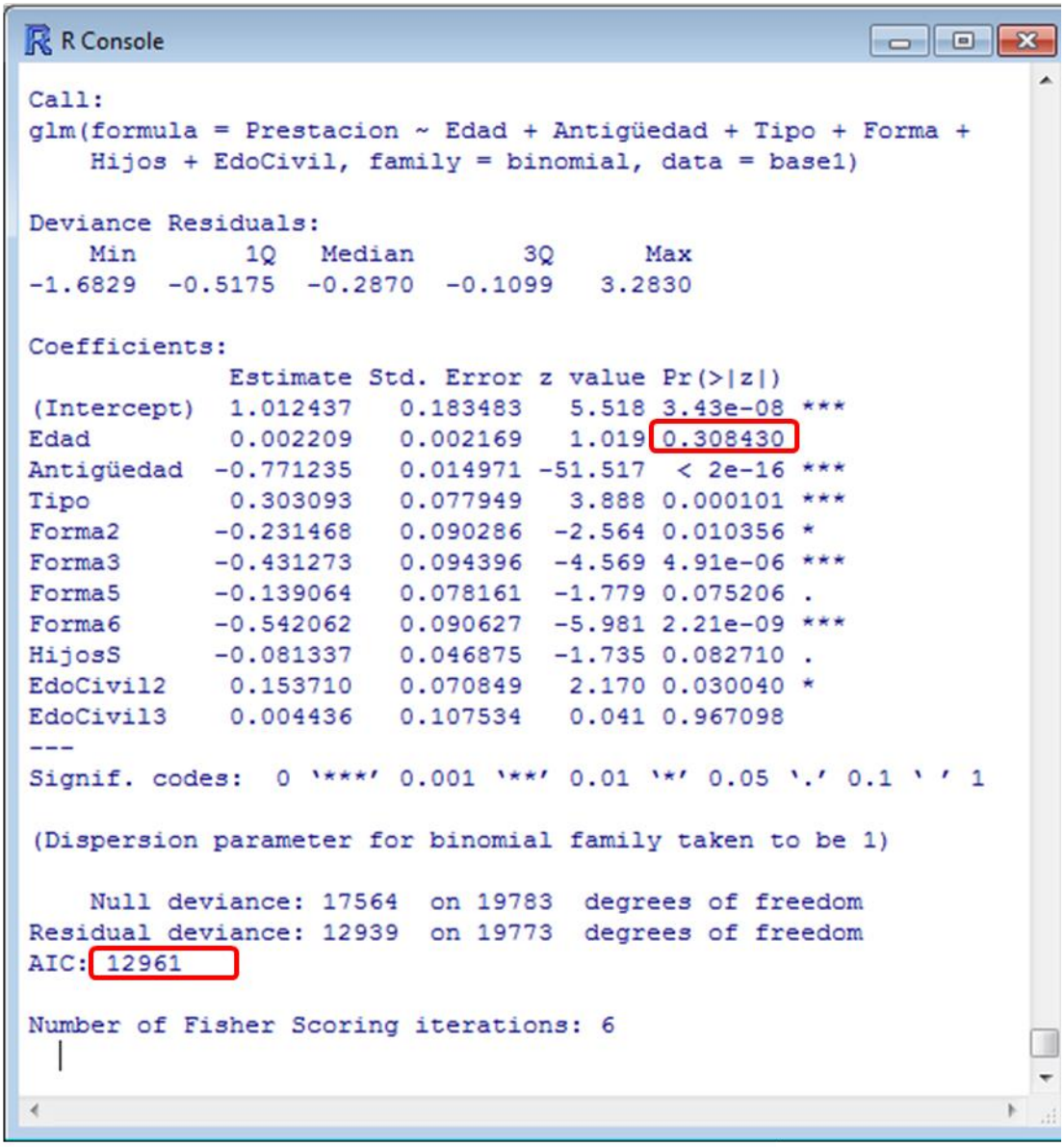


**Figura 51**: *Salida de Resultados Modelo IV – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Con este nuevo intento, no se logra bajar el nivel de AIC, sin embargo se hará un nuevo intento para verificar que se trata del AIC más bajo que se puede obtener. Con lo cual, ahora se sugiere sacar la variable EDAD, quien también muestra un p-valor por encima del nivel objetivo.

Así se obtiene el siguiente cuadro resumen de resultados *(Figura 52)*:

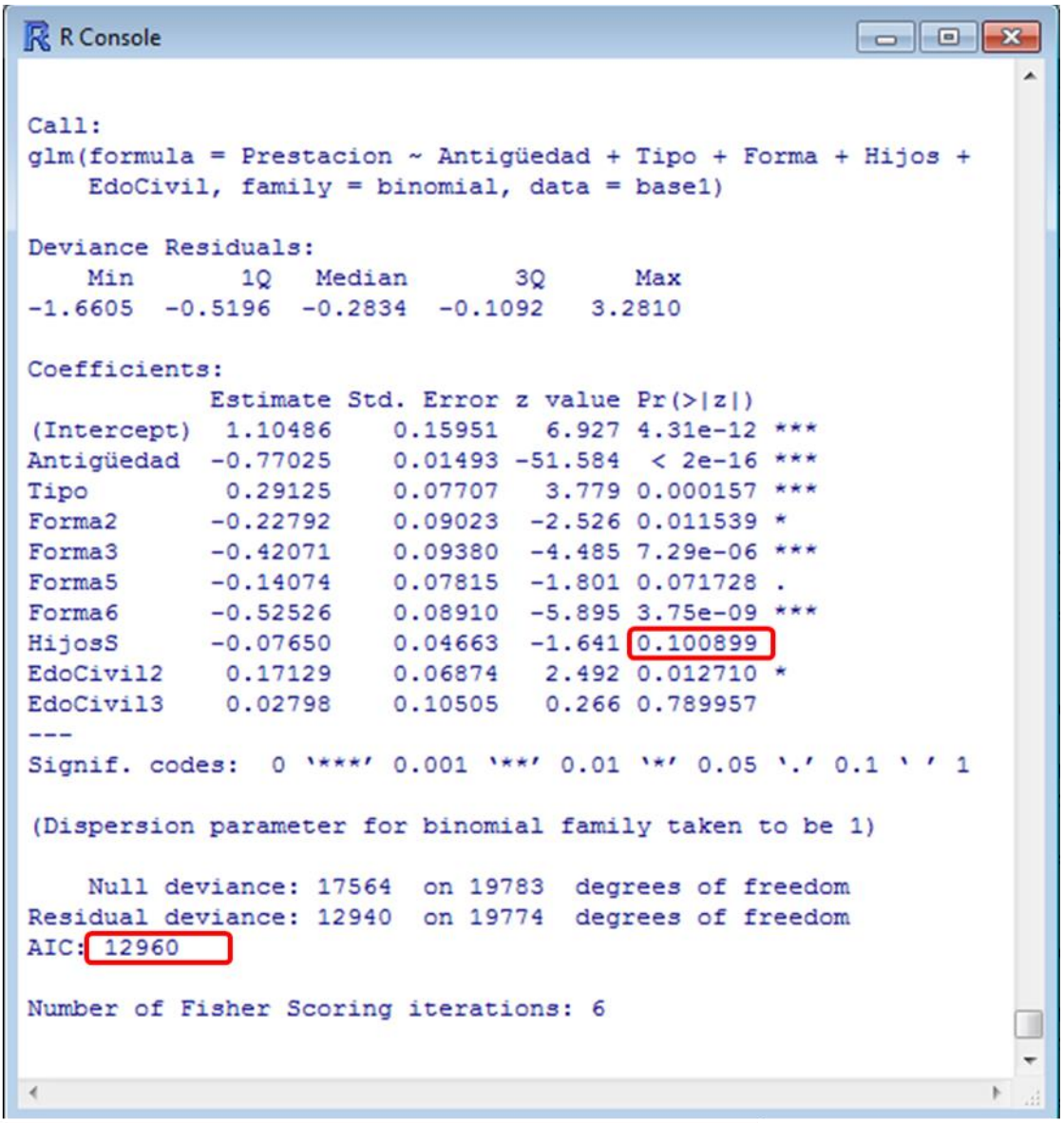


**Figura 52**: *Salida de Resultados Modelo V – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Efectivamente con este nuevo intento, se logra reducir en 1 punto el AIC, con lo cual, parece buena opción, tanto quitar del modelo la variable ICE como la variable EDAD.

Sin embargo, seguimos teniendo p-valor por encima del objetivo; esto es la variable HIJOS cuenta con un p-valor del *0.1.* Como ya se ha mencionado, esta decisión de eliminarlo o no, será dependiendo de cómo de conservadores o no se quiera ser en el momento de la modelización.

Pues bien, siguiendo la técnica que se ha venido tomado de "prueba-error"; y debido a lo relativamente sencillo que es obtener los resultados, sacando esta variable, gracias a la aplicación de R que se ha utilizado; se procede a obtener una nueva opción *(Figura 53)*:

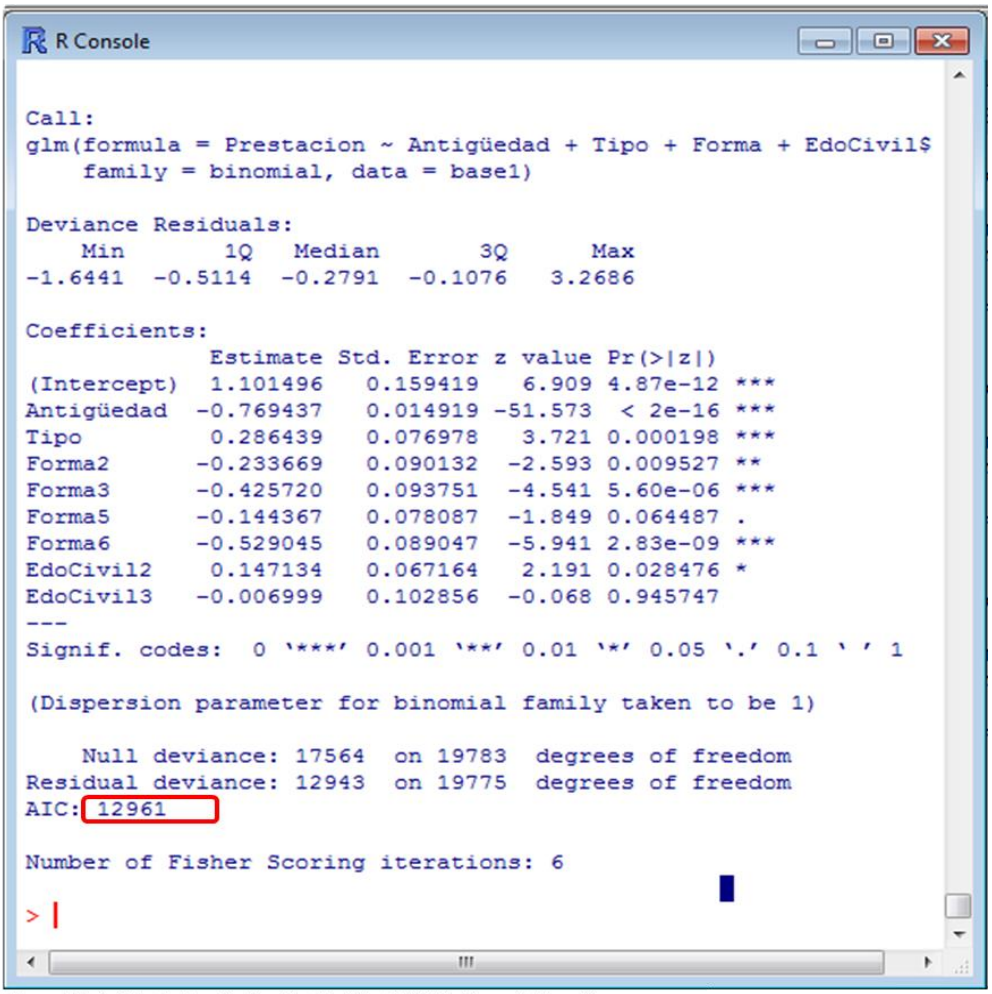


**Figura 53**: *Salida de Resultados Modelo VI – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Pues efectivamente con este sexto intento, aunque los p-valores logran el objetivo del 95% de nivel de confianza; el AIC no se reduce sino por el contrario incrementa en 1 punto. Por lo que ahora el tema estará en saber con el cual de estos 2 últimos intentos corresponderá el modelo óptimo. Es así como se recurre a la metodología de ANOVA para comprobar la significación de los factores incluidos en el modelo.

Los resultados del ANOVA para el primer modelo serian *(Figura 54)*:

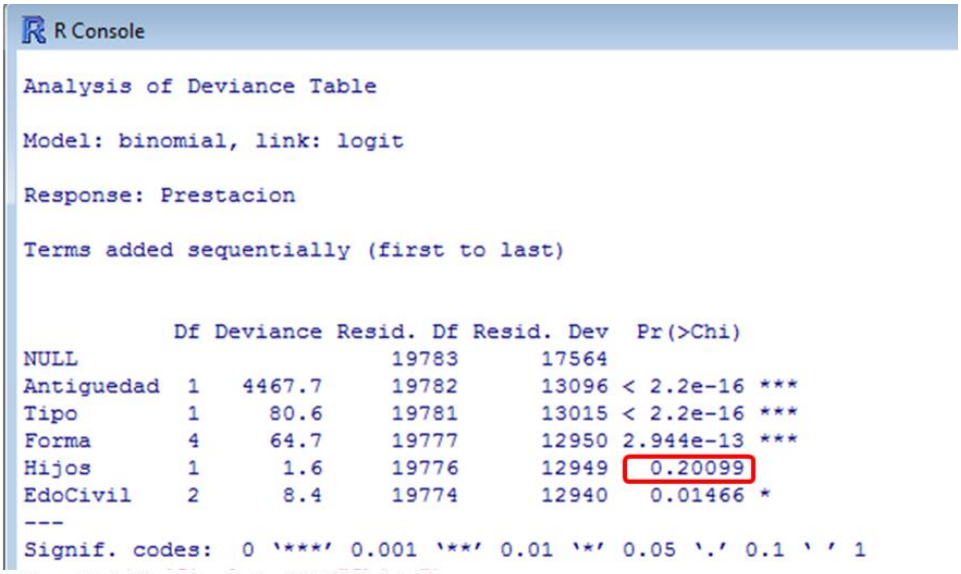


**Figura 54***: Salida de Resultados ANOVA Modelo V*
**Fuente**: *Propia de los autores*

Y finalmente el mismo ejercicio ANOVA para la segunda opción modelo, se tiene *(Figura 55)*:

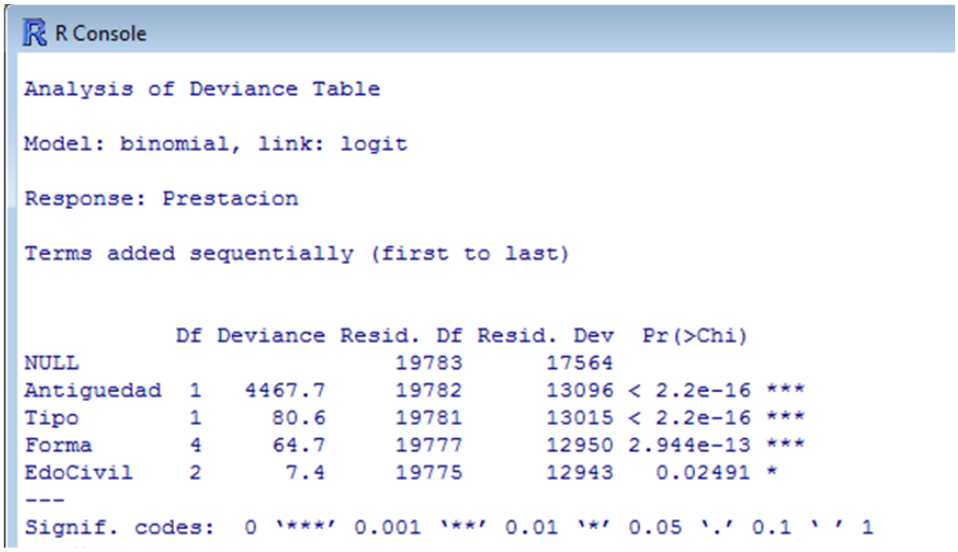


**Figura 55***: Salida de Resultados ANOVA Modelo VI*
**Fuente**: *Propia de los autores*

Se observa que en esta segunda opción, todos los p-valores se encuentran por debajo del nivel objetivo; es decir menor que el 0.05. Lo indicaría que se trata de un mejor modelo. Por lo que, a pesar que se podría decidir sacrificar el punto que se ganó en el AIC, pero se gana en ajuste, logrando algo por arriba del 95% inclusive.

#### 5.4.2.3. Diagnóstico del Modelo

Para hacer el diagnóstico del modelo, se considerará el último modelo, ya que es el que mejor resultados obtuvo, tanto en tomando el criterio AIC como los p-valores resultantes del ANOVA aplicado.

Pues bien, mediante la Devianza, se puede conocer el porcentaje de probabilidad de caída de cartera podría ser explicado por el modelo:

$$D^2 = \frac{DevianzaModeloNulo - DevianzaResidual}{DevianzaModeloNulo} * 100$$

Para el caso del modelo sería:

$$D^2 = \frac{17564 - 12943}{17564} * 100$$

$$D^2 = 26.31\ \%$$

Ahora bien, se hace un análisis visual de los resultados, se puede recurrir a la Curva ROC (siglas en inglés *"Receiver Operating Characteristic"*). Se trata de una Teoría de detección de señales donde mediante la representación gráfica del ratio de "Verdaderos Positivos" frente al ratio de "Falsos Positivos". El análisis de la curva ROC proporciona una herramienta para seleccionar los modelos posiblemente óptimos. Se recurre a ella como una medida para la elección entre pruebas diagnósticas distintas.

En la siguiente figura se muestran algunas de las diferentes curvas de ROC que podrían resultar *(Figura 56)*:

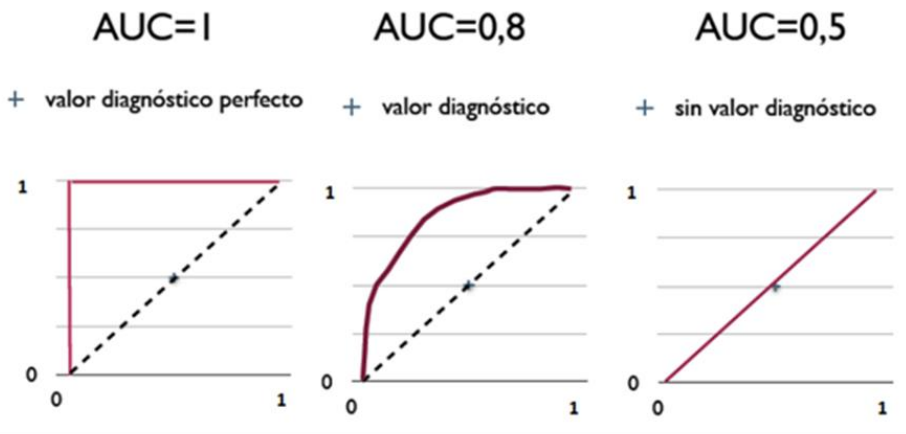


**Figura 56***: Diferentes tipos de Curvas ROC*
**Fuente**: *https://commons.wikimedia.org/wiki/File:Curvas.png*

La forma de interpretar el resultado de la prueba sería observando el área bajo la curva en ambas pruebas. Su valor está comprendido entre 0.5 y 1; donde 1 representa un valor diagnóstico perfecto de la prueba; y 0.5 es una prueba sin capacidad discriminatoria diagnóstica. A continuación la curva ROC resultante del modelo seleccionado sería *(Figura 57)*:

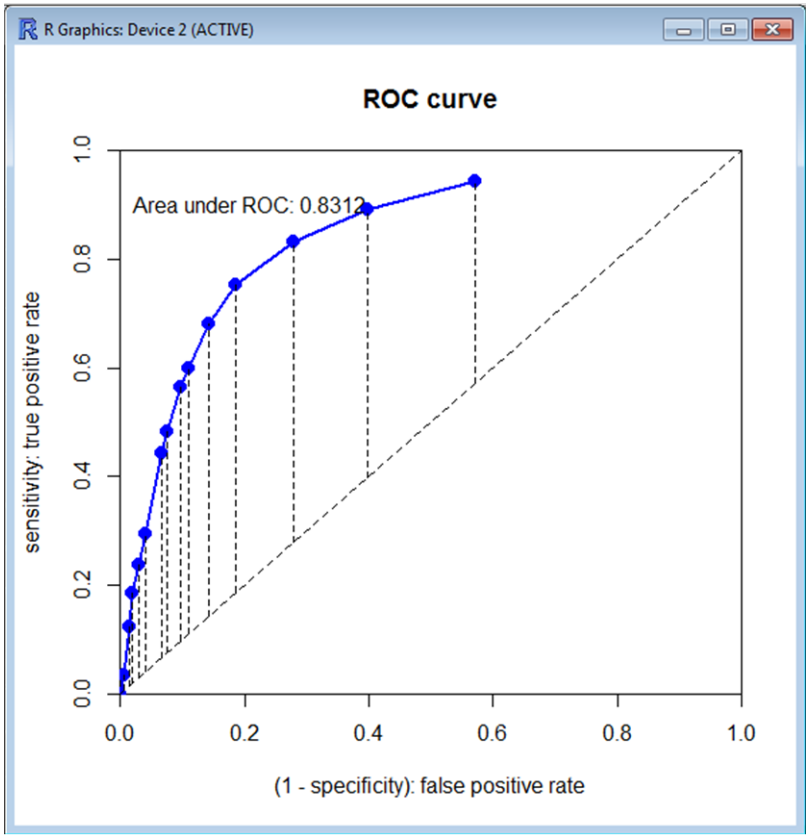


**Figura 57***: Curva de ROC del Modelo VI – R (GLM, Binomial, Logit)*
**Fuente**: *Propia de los autores*

Es decir, el resultado arroja un 83,12% lo cual significa que se trata de un Test bueno, ya que supone que la probabilidad de "diagnosticar" a una póliza como candidata a ser anulada sea correctamente clasificada es del 83,12%. Únicamente, a manera de contraste, se ha generado la curva ROC para los otros modelos anteriores al seleccionado; y en todos los casos, éste último es el que resulta con el mejor porcentaje, lo cual confirma que la elección del modelo ha sido acertada.

### 5.4.3. Principales Resultados Obtenidos bajo GLM

Una vez seguido, paso a paso, las fases propuestas para la implementación de un Modelo Lineal Generalizado; se pueden resumir los principales resultados obtenidos e identificar las implicaciones prácticas que pudiesen tener dichos resultados.

Uno de las opciones que ofrece la aplicación de un Modelo Lineal Generalizado, para el análisis de los resultados, es la interpretación gráfica de los factores por variable. Esto es, cual es el efecto que tiene los niveles de cada variable con respecto a su Nivel Base *(Figura 58)*:

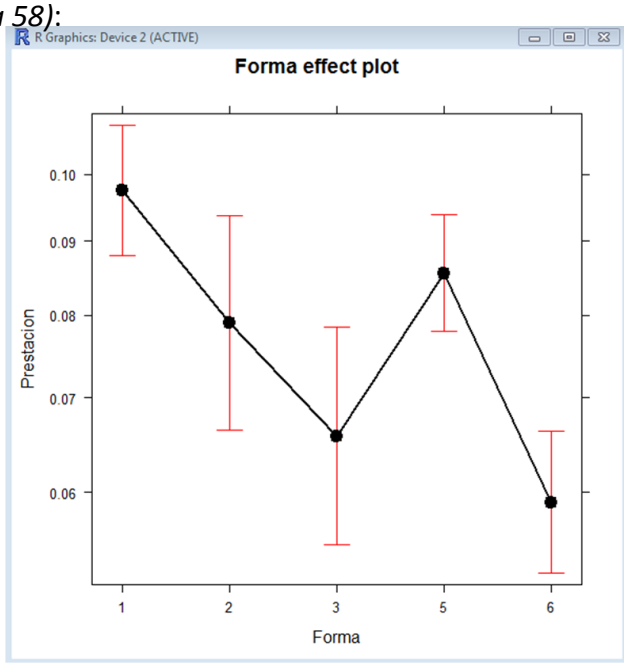


**Figura 58**: *Efecto de la variable FORMA PAGO con respecto a su Nivel Base*
**Fuente**: *Propia de los autores*

Esto se interpreta, tomando como Nivel Base la FORMA PAGO=1 (Anual), el resto de niveles que toma la variable, llevarían un efecto negativo. Esto, en otras palabras, tomando como base las pólizas con Forma de Pago Anual, la propensión a la cancelación de la póliza, incrementa conforme incrementa el valor de la categoría. Sabiendo que el valor FORMA PAGO=2 (Semestral), querría decir que este tipo de pólizas es mayormente susceptible a anular su contrato de seguros; que las pólizas anuales. Lo mimos para la categoría 3 (Trimestral) y así sucesivamente.

Por otro lado, otra de los resultados que se pueden obtener de la aplicación de un Modelo Lineal Generalizado, se pueden mostrar la capacidad predictiva de este tipo de modelos mediante el siguiente gráfico *(Figura 59)*:

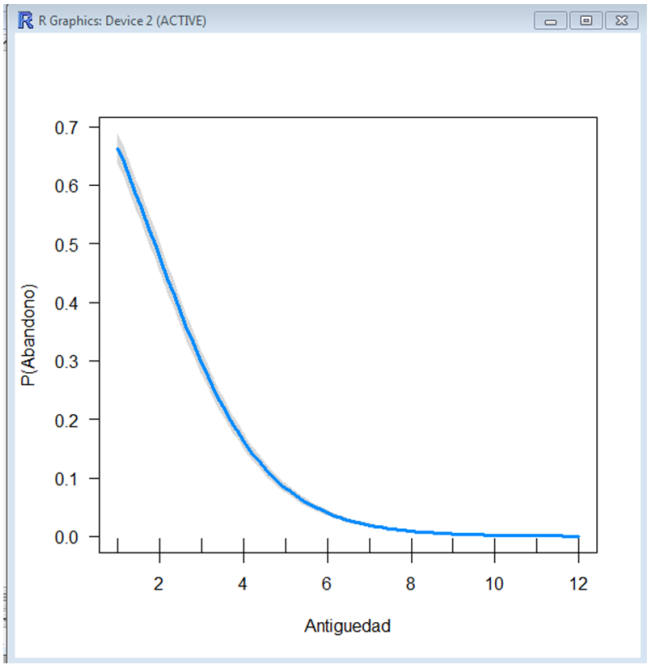


**Figura 59:** *Probabilidad de Caída o Abandono vs Antigüedad de la Póliza*
**Fuente:** *Propia de los autores*

Lo que se puede observar en el gráfico es que entre más años de Antigüedad se tiene en la compañía de seguros, decrece la probabilidad de anular la póliza que se tiene contratada. Lo cual suena lógico, por la fidelización de la entidad hacia sus clientes.

Sin embargo, el objetivo del presente estudio no es llegar a una probabilidad de abandono o cancelación, buscando una fiabilidad absoluta de la respuesta predictiva del modelo. Sino; retomando el objetivo inicial, es obtener el conjunto óptimo de factores o variables que definen el perfil del asegurado susceptible a la cancelación de su póliza.

Por lo que, resumiendo los resultados obtenidos de la aplicación empírica de un Modelo Lineal Generalizado, se ha obtenido el modelo óptimo cuyo conjunto de factores seleccionados como variables explicativas serían: ANTIGÜEDAD, TIPO PRODUCTO, FORMA PAGO, EDO CIVIL.

Ahora bien, contrastando estos resultados con algunos análisis que se han realizado en el estudio de pólizas caídas en el ramo de No Vida, se tiene el efecto de la Edad del tomador de la póliza de seguros *(Figura 60)*:

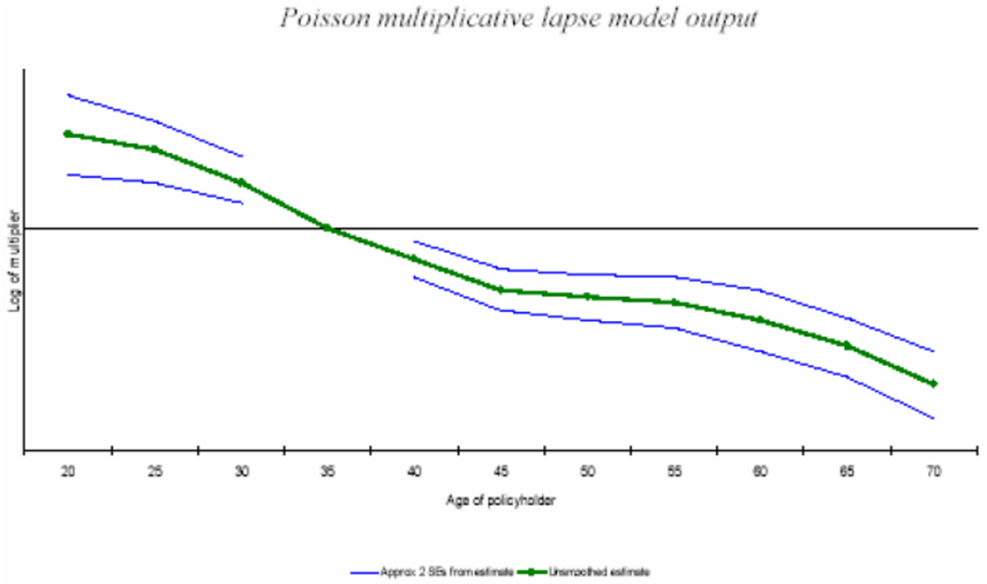


**Figura 60**: *Efecto de la variable EDAD dentro de un GLM para el Ramo de No Vida*
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Se observa que los aseguradores jóvenes tienen una mayor tasa de anulación que los mayores, probablemente porque tienen más tiempo libre y entusiasmo a la hora de buscar mejores condiciones; o bien simplemente porque no se encuentra tan desarrollado el interés asegurable y calidad del producto y sólo buscan un precio competitivo.

Por otro lado, se puede recurrir a los resultados obtenidos en un caso real en que se ha empleado para analizar las caídas y rescates sufridas por una aseguradora italiana. En este estudio se concluyó que los factores de riesgo de anulación son: Producto, Año de Exposición, Duración y Año de Suscripción de la Póliza. Con lo cual, se podría tener como factor coincidente la ANTIGÜEDAD de la póliza y el TIPO PRODUCTO.

Tomando las conclusiones del análisis de los efectos de la duración de la póliza, se comentan que aquellos asegurados que no han rescatado en los primeros 10 años, difícilmente lo harán después (*Figura 61*)

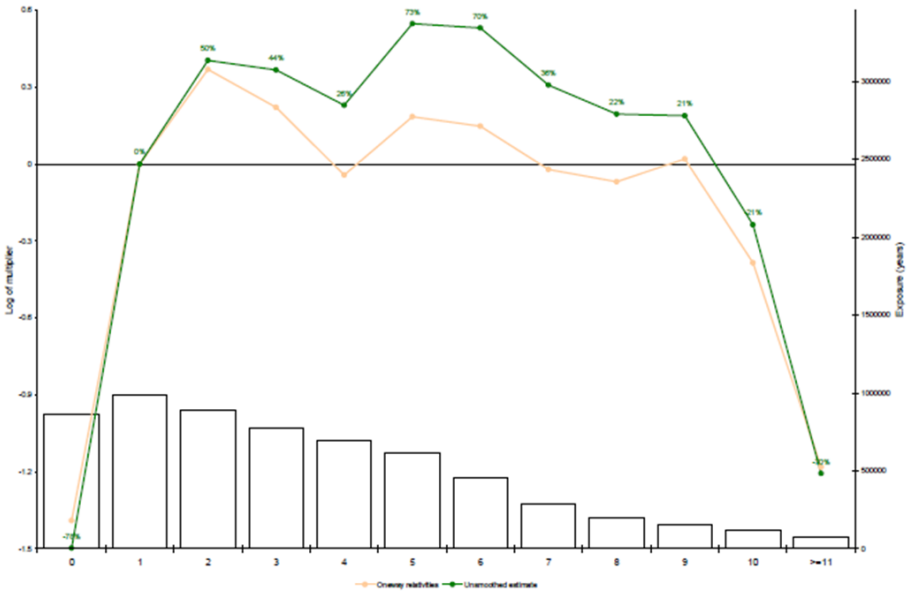


**Figura 61**: *Efecto de la variable DURACION dentro de un GLM para el Ramo de No Vida*
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

En cuanto al tipo de Producto, se habla de también tener un fuerte efecto; aunque en este caso , no podría ser cien por ciento comparable, ya que mucho dependerá del tipo de cartera que se tiene (*Figura 62*):

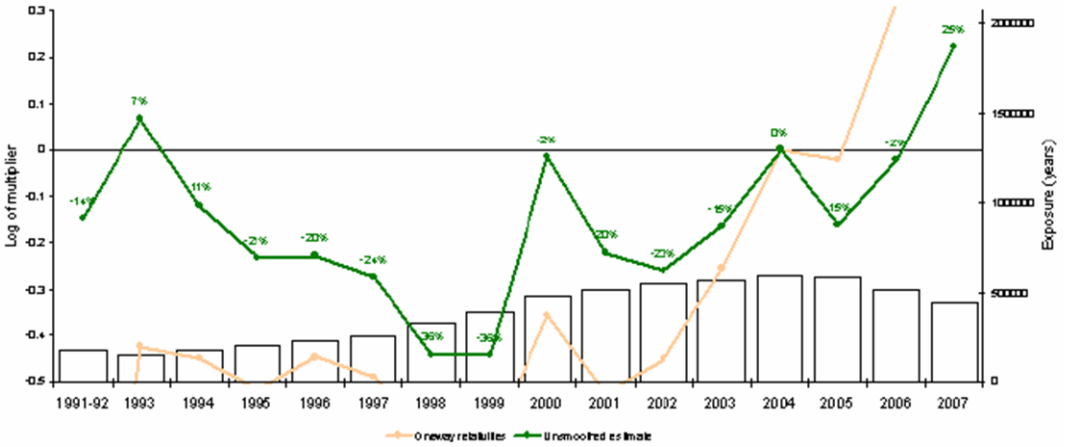


**Figura 62**: *Efecto de la variable PRODUCTO dentro de un GLM para el Ramo de No Vida*
**Fuente**: Propia de los autores a partir de: *Duncan Anderson, Sholom Feldblum. "A Practitioner's Guide to Generalized Linear Models". Feb 2007*

Con lo cual, el modelo óptimo que arroja las variables ANTIGÜEDAD, TIPO PRODUCTO, FORMA PAGO y EDO CIVIL; se podría coincidir en 2 de ellas con el caso italiano, y por tanto, pudiendo aportando 2 factores más a estudiar en una cartera real dentro de una entidad aseguradora.

Así, a manera de resumen general, se comparan estos resultados obtenidos del modelo paramétrico con respecto a las variables significativas que se han obtenido de las técnicas no paramétricas de la Inteligencia Artificial; para cada una de las categorías o clases analizadas: Categoría 1- CAIDA (*Tabla 38*) y Categoría 0 – RETENCION (*Tabla 39*):

**COMPARATIVO DE METODOLOGÍAS**
**RESUMEN DE VARIABLES SIGNIFICATIVAS**

| **Categoría: 1** |
|---|
| **CAIDA** |

| **NO PARAMETRICO** | | **PARAMETRICO** |
|---|---|---|
| **ARBOLES DE DECISION** | **ROUGH SET** | **GLM** |
| Antigüedad | Antigüedad | Antigüedad |
| Tipo Producto | Forma Pago | Tipo Producto |
| Edad | Tipo Producto | Forma Pago |
| Edo Civil | Edo Civil | Edo Civil |
| Forma Pago | Hijos | |
| Sexo | | |
| Hijos | | |
| ICE | | |
| Nivel Estudios | | |
| Nivel Ingresos | | |

**Tabla 38:** *Comparativo de Resultados de Metodologías – CLASE 1: CAIDA*
**Fuente**: *Propia de los autores*

Para la CLASE 1 correspondiente a las variables o patrones de comportamiento que definen a los clientes susceptibles a la anulación de su contrato de seguros, se obtienen las mismas variables significativas que arroja el modelo GLM como factores con mayor significancia.

Ahora bien, para la CLASE 0 correspondiente a los clientes propensos a quedarse en la compañía conservando su póliza contratada, sucede algo similar; aunque en diferente orden de aparición pero continuando siendo de las variables más significativas para clasificar a los clientes *(Tabla 39)*:

## COMPARATIVO DE METODOLOGÍAS
## RESUMEN DE VARIABLES SIGNIFICATIVAS

| **Categoría: 0** | | |
|---|---|---|
| **RETENCION** | | |
| | | |
| **NO PARAMETRICO** | | **PARAMETRICO** |
| **ARBOLES DE DECISION** | **ROUGH SET** | **GLM** |
| Antigüedad | Antigüedad | Antigüedad |
| Edo Civil | Tipo Producto | Tipo Producto |
| Tipo Producto | Forma Pago | Forma Pago |
| Edad | | Edo Civil |
| Red | | |
| Forma Pago | | |
| Sexo | | |
| Nivel Estudios | | |
| Nivel Ingresos | | |

**Tabla 39:** *Comparativo de Resultados de Metodologías – CLASE 0: RETENCION*
**Fuente**: *Propia de los autores*

## CAPITULO 6: CONCLUSIONES GENERALES

Como se ha comentado, existen varias causas y finalidades que han provocado el surgimiento, desarrollo y lanzamiento del cercano marco normativo sobre el que descansará el sector asegurador mejor conocido como Solvencia II.

Sin embargo, esta eminente e inesperada evolución del proyecto, hace que se empiece a olvidar la filosofía básica de Solvencia II; derivando en una serie de normas que haga imposible en la práctica, la oferta de productos y servicios que son posibles en el presente del seguro europeo. En otras palabras, si el sector asegurador ha alcanzado niveles de capitalización financiera adecuadamente sólidos, es evidente que, por mucho que se deban ajustar los perfiles de riesgo, no se deberían pensar en encontrar sorpresas inesperadas por parte de la actual gestión de la cartera. Si por consiguiente, esta nueva normativa derivase en el abandono por parte de las entidades aseguradoras de cierta modalidad de productos, para lograr cumplir con una visión correcta de la solvencia del sector, sería un fallo hacia la demanda de dichos productos, es decir un fallo hacia los clientes; olvidándose de que la razón de existencia de toda oferta es que exista una demanda que la necesita.

Ante esta situación, el proyecto se ve paralizado al intentar dar respuesta a la cuestión de si todos estos objetivos son compatibles. La creación de Solvencia II es un proceso complicado desde un punto de vista de la coordinación dentro del entorno político-económico que asume cada país miembro del sistema y la existencia de los diversos períodos de transición que enfrenta cada uno de ellos.

Como se ha intentado transmitir, Solvencia II es un proyecto sumamente ambicioso ya que no sólo busca re-diseñar la actual metodología de cuantificación de la solvencia de las entidades aseguradoras; y con ello, establecer los niveles de requisitos de capital que necesitan para hacer frente a los riesgos adquiridos frente a sus asegurados. Sino que a partir de su estructura de tres pilares, aportar una nueva cultura enfocada a optimizar la gestión de riesgos dentro del sector asegurador.

Por un lado, dentro del Pilar I se busca analizar y determinar el perfil de riesgos que pretende administrar y soportar cada entidad. Esto se obtendrá en la medida en que se logre una mayor calidad en la gestión de riesgos mediante mejores técnicas de estrategia, planeación y administración; de tal forma que las compañías aseguradoras puedan ser capaces de mantener su posicionamiento frente a los diferentes riesgos que soporta mediante su identificación y control de la evolución constante de los mismos.

Ahora bien, en cuanto a la calidad en la supervisión que se plantea dentro del Pilar II, se pretende mejorar y principalmente, homogeneizar la actuación de las autoridades supervisoras dentro del ámbito europeo. Con ello, se busca implementar nuevos procedimientos de detección de situaciones de peligro con suficiente antelación, que puedan perjudicar la solidez financiera, estructural y cultural de las entidades; y por tanto puedan incurrir en un desequilibrio o amenaza del mercado asegurador.

En lo que se refiere al Pilar III, durante el proceso de preparación y realización de la información que se exige en esta sección, las entidades también pueden encontrarse con temas de deficiencias de información y establecimiento de controles para garantizar la calidad e integridad de los datos, cuya resolución se traduce en tiempo y recursos relevantes para su ejecución. Es decir, la generación de la información exigida por este pilar, son parte fundamental del proceso de implementación de Solvencia II que toda entidad aseguradora debe planificar debido a la gran dedicación y esfuerzo que se debe invertir en ello y por lo tanto no debiese dejar aislado y para el último momento.

Por lo que es evidente que será un proceso largo y que presenta importantes retos tanto cuantitativos, financieramente hablando; como cualitativos en cuanto al nuevo gobierno organizativo y cultural que exige este nuevo entorno; y todo ello bajo una total transparencia y disciplina del mercado asegurador en su totalidad. Es por ello, que Solvencia II necesita entidades aseguradoras solventes cuyas decisiones estratégicas se tomen en función de esta nueva <<cultura>> de la gestión del riesgo; sin verse por ello amenazadas ante posibles debilidades de su patrimonio y prestigio

para finalmente lograr traducirse en un inminente fortalecimiento del sector asegurador.

Por tanto, se han expuesto las implicaciones que conlleva la nueva regulación de Solvencia II dentro de la gestión de riesgos que deben asumir las entidades aseguradoras. Una de estas implicaciones es precisamente la relevancia que supone la correcta cuantificación del riesgo de cartera. Esto ha llevado a hacer un análisis minucioso del mismo. Concluido dicho análisis, es momento de recapitular y enmarcar las conclusiones a las que se han llegado a lo largo de la investigación realizada.

Retomando el objetivo que se ha planteado para este estudio, se hablaba de lograr identificar una serie de variables o patrones de conducta que caracterizan a los tipos de clientes susceptibles a la cancelación de su contrato de seguros; con la finalidad de establecer estrategias comerciales de retención de clientes en aquellas pólizas con poca propensión a la anulación de su póliza; o bien lograr una gestión eficiente de la caída de cartera y el riesgo que conlleva. En otras palabras, se traduce en utilizar la metodología que ofrece Inteligencia Artificial, contribuyendo al equilibrio y estabilidad de la solvencia que las entidades aseguradoras requieren.

Es más, estando en contexto de los niveles de exigencia que propone el entorno regulatorio, surge la necesidad de buscar metodologías de análisis novedosas; de tal forma, que se logre incluir o detectar características de los clientes con perfil de "anulador" que ayuden a complementar los métodos que hasta el día de hoy se han venido utilizando dentro del sector asegurador. De esta forma, se han recogido las principales características de la Inteligencia Artificial y se ha optado por la utilización de dos de sus técnicas como son los "Arboles de Decisión" y la "Teoría Rough Set".

Con base en ello, se ha procedido a aplicar dichas técnicas que provienen de la Inteligencia Artificial utilizando una base de datos de clientes de una entidad aseguradora. Para ello, se ha realizado un análisis exploratorio de las variables utilizadas, así como un breve resumen del contexto y características de la muestra utilizada.

De los resultados de la utilización de las dos técnicas de Inteligencia Artificial, se pueden obtener tres conclusiones principales con base en la identificación de tres tipos de comportamientos de acuerdo a la duración de sus contratos de seguro o antigüedad que tienen dentro de la compañía. Esto es, se pueden definir un sistema de diagnóstico rápido para identificar los clientes susceptibles a la anulación a: corto plazo (antigüedad de 1 o 2 años), mediano plazo (de 3 a 5 años de antigüedad) y largo plazo (de 5 años en adelante dentro de la compañía). De esta forma, poder segmentar la cartera de pólizas de la entidad, de acuerdo al grado de fidelidad que ha mostrado el cliente hacia la compañía.

Así, se tiene que para clientes de corto plazo, es decir que tienen muy poco tiempo con su póliza de seguros, es clave la Forma de Pago contratada. Los resultados obtenidos indican que, tratándose de una póliza de Vida Ahorro, la cual se paga Mensualmente, son características propias de un perfil "anulador". Esto tiene su lógica considerando que el cliente tiene la posibilidad de "auto-preguntarse" continuamente: si requiere o no, si desea o no, si considera viable o no el hecho de mantener su póliza de seguros en vigor. Es decir, en el caso contrario cuando se trata de una póliza Anual, estas preguntas se presentan únicamente una vez al año. Sin embargo, en el caso de una póliza Mensual, el cliente se cuestiona continuamente si debería seguir pagando su póliza de seguros, siendo el pago de este servicio un tema no prioritario en momentos de poca liquidez o crisis financiera familiar.

Para el caso de clientes definidos como de mediano plazo, esto es pólizas con duraciones entre 3 o 5 años, las conclusiones obtenidas se dirigen hacia el seguimiento de los clientes cuyo Estado Civil declarado ha sido Casado. Estos resultados, más que asociarlo a un comportamiento racional o deductivo, puede ser interesante en temas de gestión de cartera. En otras palabras, quizá no sea del todo preciso considerar esta regla de decisión como un patrón predictivo; ya que se estaría "discriminando" a gran parte de la población, las gente Casada, cuyo interés asegurable es significativo. Más bien, sería interesante fomentar el seguimiento y control de este tipo de cartera; es decir, implementar sistemas de alarmas basándose en estos resultados, donde se ponga especial atención a la cartera de clientes con un Estados Civil determinado, ya que pueden presentar cierta tendencia a cancelar su póliza de seguros.

Ahora bien, considerando que se trata de una cartera de clientes con más de 5 años de antigüedad, a quienes se ha considerado como pólizas de Largo Plazo, una variable que ha sido generada para "calificar" a los clientes, el Valor del Cliente, juega un papel interesante. Se ha logrado identificar que aquellos clientes clasificados por la entidad como "Vinculados y Medianamente Rentables"; tienden a cancelar su póliza al cabo de ciertos años. Hasta cierto punto, sonaría lógico que estando "Vinculados" (lo cual se mide con la cantidad de pólizas contratadas por el cliente), es decir, con más de una póliza contratada, ya se ha fidelizado al cliente y por tanto, no se espera que salga de la compañía. Pero por otro lado, consideremos la segunda frase "Medianamente Rentables", lo cual quiere decir, que el cliente cuenta con muchas pólizas sin embargo, no alcanza a ser "buen cliente" en términos de rentabilidad. Esto puede tener implicaciones interesantes para la entidad ya que podría cuestionarse si es conveniente tener este tipo de clientes. Es decir, un cliente que genera costes y sobrecarga de trabajo operativo hacia la compañía por el volumen de pólizas que tienen; sin embargo, en términos de rentabilidad no es del todo significativo; y finalmente, al cabo de unos años, terminará saliendo de la compañía. En otras palabras, la entidad deberá ser consciente en el tipo de riesgo que desea asumir, por un lado, basándose en la "calidad" de clientes que conformen su cartera (rentables o poco rentables); o bien, en la "cantidad" de pólizas de la misma (volumen de su cartera sacrificando rentabilidad).

Adicionalmente a estos resultados, se ha logrado detectar un cierto patrón de comportamiento que identificaría a los clientes que, por el contrario, presentan cierta tendencia a mantener su póliza de seguros en vigor. Ésta indica que aquellos clientes con póliza de Vida Riesgo, con Hijos y mayores a 36 años, no buscan cancelar su contrato. Este comportamiento suena razonable, considerando que las personas con este perfil, pueden tener mayor consciencia e interés en mantener este tipo de seguros. Esto sugiriendo que, una persona con mayor madurez y con hijos, busca asegurar un patrimonio ante la incertidumbre de su futuro y cuyo interés es mantener la estabilidad familiar ante cualquier eventualidad. Nuevamente, esta conclusión puede presentar implicaciones interesantes en términos de retención de cartera de

clientes, más aún en los tiempos que corren cuando hablar de crecimiento es mucho más complejo que cuidar la cartera ya conseguida a lo largo de los años.

Por otro lado, también se ha propuesto la aplicación de una técnica paramétrica con el fin de contrastar los resultados obtenidos por las técnicas no paramétricas que ofrece la Inteligencia Artificial. Estas dos disciplinas se han desarrollado en el entorno académico, una a la espalda del otro (Banet Tomás, 2001). Por un lado, la Estadística paramétrica se ha preocupado por el poder de la generalización de los resultados obtenidos para poder inferir hacia situaciones más generales que la estudiada. Por el contrario, las técnicas no paramétricas, no le interesa las distribuciones de los datos con los que se trabaja, sino que ofrecen soluciones algorítmicas con un coste computacional aceptable.

Con base en ello, se ha dedicado una sección completa a la recopilación de la teoría que existe detrás de los Modelos Lineales Generalizados. Si bien es cierto que no se expone nada nuevo técnicamente, sí se consideró importante mencionar la complejidad de su teoría y así mostrar cuáles eran las ventajas y por qué resultaría o no efectivo su aplicación en el ramo de Vida. Así mismo, recabar la información sobre la estructura, componentes y parámetros de los GLM, es de gran utilidad para poder argumentar a favor de su implementación en el ramo de Vida. Es así como otro de los bloques fue dedicado a reunir información que se debe tener en cuenta sobre aplicación práctica de un Modelo como éste; ya que existen varios procesos que deben ser tomados en cuenta en su desarrollo con el fin de obtener los mejores resultados en la medida de lo posible.

Con toda esta información se procedió a la aplicación empírica de un GLM sobre la misma base de datos con la que se trabajó las técnicas de Inteligencia Artificial. Tras los resultados obtenidos, primeramente se puede confirmar que es posible la aplicación de este tipo de metodologías en el análisis del riesgo de caída de cartera. Ahora bien, a pesar de tener cierta complejidad el tema de la interpretación de los resultados “puros” que van arrojando las iteraciones del modelo; se ha podido llegar a un modelo con una bondad de ajuste bastante satisfactoria. Es por ello, que el objetivo sólo fue obtener factores o variables que proporcionarán el perfil del

asegurado “anulador” o susceptible a cancelar su contrato de seguros. Y de esta forma, dicha información compararla con los resultados anteriores.

Por tanto, recapitulando los resultados, ambas técnicas sugieren que las variables que se deben tener en cuenta como posibles patrones de comportamiento son:

- Antigüedad de la Póliza
- Tipo de Producto a la que pertenece la Póliza
- Forma de Pago de la Póliza
- Estado Civil del Asegurado

Ahora bien, no se puede dejar de mencionar que las técnicas de Inteligencia Artificial, sugieren otras variables como son: Si se tiene o no Hijos, incluso el Sexo y Edad del Asegurado. Sin embargo, estas no se obtienen como variables explicativas significativas en el modelo GLM.

Así mismo, mencionar que al tener una variable respuesta dicotómica: Cancela o Renueva; es posible estudiar el efecto contrario a la anulación del contrato de seguros. Es decir, se podría plantear el objetivo de manera inversa y analizar el patrón de comportamiento del perfil del asegurado fiel a la entidad. De aquí la propuesta de utilizar ambas metodologías para temas de retención o conservación de clientes.

Por último, se debe suponer que exista la opinión de las entidades aseguradoras de que no se está avanzando en nada nuevo; ya que muchas de las conclusiones parecen ya sabidas por el sector. Sin embargo, se considera que la mayor aportación de este trabajo es la de incentivar al sector a buscar otro tipo de técnicas para la gestión de sus riesgos. Entre líneas, se debe leer que Solvencia II es una “cultura de cambio”; y como parte de esa renovación, surge el planteamiento de este nuevo tipo de técnicas. Si bien, no buscando que éstas suplanten a las ya tradicionales técnicas estadísticas, si podrían complementar el sentido común de los expertos dedicados a modelos de gestión de riesgos dentro de las entidades. De alguna manera darían un enfoque distinto y complementario a los tradicionales modelos de gestión de riesgos usados por el sector.

No es por demás comentar que, este estudio no está exento de limitaciones. Por un lado, el tema de la muestra se ha limitado a los productos con mayor volumen de producción y sólo se ha considerado un año de ejercicio contable de la entidad; con lo cual se trata de una muestra limitada en cuanto al número de casos analizados. Por otro lado, el número de variables cualitativas analizadas también se encuentra limitado, ya que las entidades aseguradoras han tenido poco interés en capturar en sus bases de datos, demasiada información cualitativa del asegurado. Es hasta hace algunos años, cuando se empezó a hablar de la "Calidad de los Datos" a raíz de las exigencias marcadas por Solvencia II en el tema.

Finalmente, es necesario recalcar que lo que se perseguía con este estudio no era tanto la bonanza, fiabilidad y precisión de los resultados; sino presentar, aplicar y discutir la factibilidad y capacidad de aplicar las metodologías ofrecidas por la teoría de Inteligencia Artificial en el campo de los Seguros de Vida. Enlazando esto con lo anterior, los resultados pueden variar en la medida en la que se logre la mayor calidad de la información utilizada. Lo mismo sucede con los resultados obtenidos del modelo GLM; ya que de hecho no se buscaba obtener la ecuación multivariada final con la que se podría predecir una tasa de caída o anulación (o bien tasa de retención). No por ello, se debe dejar de mencionar el poder predictivo de este tipo de modelos, aun cuando también se debe tener en cuenta la visión de negocio o conocimiento del tipo de seguro que se está trabajando, para la toma de decisiones basada en el modelo.

Sin embargo, no dejando de considerar dichas limitaciones, se puede afirmar que se ha cumplido con el objetivo marcado inicialmente: por un lado, logrando aplicar una nueva metodología e identificando perfiles o patrones de comportamiento de los clientes susceptibles a anular su póliza de seguros; y por el otro, se ha puesto de manifiesto la conveniencia del uso de los GLM en el campo de la estadística actuarial.

Así mismo, dejando un tema interesante por abordar para las entidades aseguradoras, se busca fomentar su uso como una alternativa novedosa o bien como un complemento a los actuales métodos utilizados en el sector, para que de manera conjunta, ofrezcan una gestión oportuna de los riesgos ante la propuesta de la nueva regulación de Solvencia II.

**BIBLIOGRAFIA**


**Alcañiz Zanón, Manuela y Pérez-Marín, Ana M.** *Estrategias Innovadoras en Tiempos de Crisis*. El Sector Asegurador ante las Transformaciones del Estado de Bienestar. Fundación de Estudios Financieros. Disponible en: *www.fef.es/new/.../751_1cc6606e829420dd54e627c7b81133f4.html*

**Alegre Escolano, Antonio, Ayuso Gutiérrez, Mercedes, Guillén Estany, M., Monteverde Verdenelli, M. y Pociello García, E.** *Tasa de dependencia de la población española no institucionalizada y criterios de valoración de la severidad*. Revista Española de Salud Pública N° 79.3. Año 2005.

**Alonso, Alberto A.** *Solvencia II para Aseguradores No-Vida*. Septiembre-Diciembre 2008. Revista de Gerencia de Riesgos y Seguros n° 102 de la Fundación Mapfre. Disponible en: *www.mapfre.com/fundacion/html/revistas/*

**Araya, Roberto**. *Induction of decision trees when examples are described with noisy measurements and with fuzzy class membership.* INRIA Seminar, Projet CLOREC. Año 1994.

**Ayuso Gutiérrez, Mercedes. Guillén Estany, Montserrat. Pérez-Marín, Ana M.** *Modelos Internos en Solvencia II: Su aplicación al cálculo del coeficiente de caída de cartera*. Septiembre-Diciembre 2012. Revista de Gerencia de Riesgos y Seguros n° 112 de la Fundación Mapfre. Disponible en: *www.mapfre.com/fundacion/html/revistas/*

**Ayuso Gutiérrez, Mercedes, Guillén Estany, Montserrat. Pérez-Marín, Ana M.** *Metodología para el cálculo de caída de cartera en Solvencia II en presencia de contagio entre cancelaciones*. Anales del Instituto de Actuarios Españoles 2011. Disponible en: *www.actuarios.org*

**Banet, Tomás Aluja.** *La minería de datos, entre la estadística y la inteligencia artificial***.** Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa Nº 3,

Vol.25. Universitat Polit`ecnica de Catalunya. Año 2001. Disponible en*:* *http://dmle.cindoc.csic.es/pdf/QUESTIIO_2001_25_03_04.pdf*

**Barros Hernández, Rafael, Martínez María Isabel y Torre-Enciso.** *La nueva regulación europea de seguros privados: Solvencia II*. Boletín de Estudios Económicos. Vol°65.199. Año 2010.

**Betegón Sánchez, Leonor**. *Convergencia entre el mercado de seguros y el mercado de capitales*. 2005. Anuario Jurídico y Económico Escurialense. Real Centro Universitario "María Cristina". Época II, n° 38. Disponible en*:* *www.rcumariacristina.com*

**Blasco, Ignacio y Azpeitia, Fernando.** *Pilar III de Solvencia II: Un esfuerzo adicional*. Primavera 2013. Revista de Actuarios n° 32 del Instituto de Actuarios Españoles. Disponible en: *www.actuarios.org*

**Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A.** *Classification and Regression Trees*. Proceedings of the Thirteenth International Conference, Bari, Italy. 1996.

**Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J. y Perez-Marin, A. M.** *Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection?*. Journal of Risk and Insurance Nº 75.3. Año 2008.

**Camacho Miñano, Maria del Mar y Segovia Vargas, María Jesús.** *¿ Qué indicadores económico-financieros podrían condicionar la decisión del experto independiente sobre la supervivencia de una empresa en su Fase Preconcursal? Evidencia Empírica.* Información Financiera y Concurso de Acreedores, de la Universidad Complutense de Madrid. Cuadernos Contabilidad N° 13.32. Bogotá 2012.

**Cerchiara, Rocco Roberto, Matthew Edwards, and Alessandra Gambini.** *Generalized linear models in life insurance: decrements and risk factor analysis under Solvency II.* 18th International AFIR Colloquium. Año 2008.

**Cooley, Steven.** *Loyalty strategy development using applied member-cohort segmentation.* Journal of Consumer Marketing Nº 19.7. Año 2002.

**Comisión de las Comunidades Europeas**. *Directiva del Parlamento Europeo y del Consejo.* Febrero 2008. Propuesta modificada y presentada sobre la actividad de seguro y reaseguro y su ejercicio (Solvencia II). *COM/2008/0119 final – COD 2007/0143*. Disponible en: http://ec.europa.eu/finance/insurance/solvency/solvency2/index_en.htm

**Comisión de las Comunidades Europeas**. *Comunicación de la Comisión.* Noviembre 2007. Revisión del proceso Lamfalussy. *COM/2007/727 final – COD 55 de 28.2.2008*. Disponible en: http://eur-lex.europa.eu/legal-content/ES/TXT/?uri=URISERV:l32056

**Crosby, Lawrence A. y Stephens, Nancy.** *Effects of relationship marketing on satisfaction, retention, and prices in the life insurance industry*. Journal of Marketing Research. Año 1987.

**De Jong, Piet y Gillian Z. Heller.** *Generalized linear models for insurance data.* Vol. 136. Cambridge University Press. Cambridge 2008.

**Díaz Martínez, Z., Fernández Menéndez, J., y Segovia Vargas, M. J.** *Sistemas de inducción de reglas y árboles de decisión aplicados a la predicción de insolvencias en empresas aseguradoras*. Universidad Complutense de Madrid. XII Jornadas de ASEPUMA. 2004. Disponible en: http://www.uv.es/asepuma/XII/comunica/diaz_fernandez_segovia.pdf

**Díaz Martínez, Z., Fernández Menéndez, J., Heras Martínez, A., Del Pozo García, E. y Vilar Zanón, José Luis.** *Modelos Aditivos Generalizados aplicados al análisis de la probabilidad de siniestro en el seguro del automóvil.* Ministerio de Ciencia e Innovación de España y Universidad Complutense de Madrid. Año 2010.

**Dobson, A.J.** *An Introduction to Generalized Linear Models*. CHAPMAN & HALL/CRC. Boca Raton, Florida 2002.

**Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher y Neeza Thandi.** *A practioner's Guide to Generalized Linear Models*. Febrero 2007. Tercera Edición

**España. Ley 20/2015, de 14 de julio**, de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras. *Boletín Oficial del Estado*, 15 de julio de 2015, Núm. 168. Disponible en: *http://www.boe.es/buscar/pdf/2015/BOE-A-2015-7897-consolidado.pdf*

**España. Resolución de 16 de junio de 2014**, Dirección General de Seguros y Fondos de Pensiones. *Boletín Oficial del Estado*, 4 de julio de 2014, Núm. 162, Sec. III. Pág. 52505. Disponible en: *http://www.boe.es/boe/dias/2014/07/04/pdfs/BOE-A-2014-7049.pdf*

**Esquerda, A., Trujillano, J., Lopez de Ullibarri, I., Bielsa, S., Madronero, A. B., & Porcel, J. M.** *Classification tree analysis for the discrimination of pleural exudates and transudates*. Clinical Chemical Laboratory Medicine Nº 45.1. Año 2007.

**Fahrmair L., Tutz G.** *Multivariate statistical modelling based on generalized linear models,* Springer Verlag. New York 1996.

**Faraway, J.J.** *Extending the Lineal Model with R.* CHAPMAN & HALL/CRC. Boca Raton, Florida 2006.

**Fernández Palacios, Juan.** *Tendencias del Seguro de Vida*. El Sector Asegurador y de los Planes y Fondos de Pensiones. Revistas ICE (Información Comercial Española). Noviembre-Diciembre 2006. Disponible en: *www.revistasice.com/*

**Ferri, A., Rodríguez, P. y Romero, M. J.** *La gestión de riesgos. Estudio sobre el sector asegurador en España 2010: los aspectos cualitativos de Solvencia II*. Dir. P. Blanco-Morales y M. Guillén. Fundación de Estudios Financieros Nº 28. Año 2010.

**Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L**. *Weka. Data Mining and Knowledge Discovery Handbook*. Primavera, USA, 2005.

**Galipienso, María Isabel Alfonso, Quevedo, M. A. C., Pardo, O. C., Ruiz, F. E. y Ortega, M. A. L.** *Inteligencia artificial: modelos, técnicas y áreas de aplicación.* Editorial Paraninfo. Año 2003.

**González de Frutos, Pilar.** *El seguro español y Solvencia II: Tres conceptos básicos*. Primavera 2013. Revista de Actuarios n° 32 del Instituto de Actuarios Españoles. Disponible en: *www.actuarios.org*

**Greco, Salvatore, Benedetto Matarazzo and Roman Slowinski.** *A new rough set approach to multicriteria and multiattribute classification*. Rough sets and current trends in computing. Springer Berlin Heidelberg, Año 1998.

**Guillén Montserrat, Pérez Marín, A.M. y Nielsen, J.P.** *La duración de distintos contratos de seguros en los hogares. Un enfoque integrado*. Septiembre-Diciembre 2006. Revista de Gerencia de Riesgos y Seguros n° 96 de la Fundación Mapfre. Disponible en: *www.mapfre.com/fundacion/html/revistas/*

**Guillén Montserrat, Pérez Marín, A.M. y Nielsen, J.P.** *The need of monitoring customer loyalty and business risk in the European insurance industry*. Geneva Papers on Risk and Insurance – Issues and Practice N° 33. Año 2008.

**Haberman, S. y Renshaw, A.E.** *Generalized Linear Models and Actuarial Science.* Journal of the Royal Statistical Society Vol 45.4. Año 1996.

**Hammond, J. D., Houston, David B., y Melander Eugene R.** *Determinants of household life insurance premium expenditures: An empirical investigation.* 1967. Journal of Risk and Insurance.

**Hardin J. y Hilbe J.** *Generalized Linear Models and Extensions,* Stata Press. Año 2001.

**Hastie T., Tibshirani, R y Friedman, J.** *The elements of statistical learning*. SPRINGER. New York 2008.

**Heller, Gillian Z. y De Jong, Piet**. *Generalized Models for Insurance Data*. Cambridge University Press. New York 2008.

**Herrera, F., Hervas, C., Otero, J., & Sánchez, L.** *Un estudio empírico preliminar sobre los tests estadísticos más habituales en el aprendizaje automático*. Tendencias de la Minería de Datos en España, Red Española de Minería de Datos y Aprendizaje. Año 2004. Disponible en: *http://www.lsi.us.es/~riquelme/red/Capitulos/LMD35.pdf*

**Hernández José, Ramírez Mª José y Ferri César.** *Introducción a la Minería de Datos.* Pearson Educación. Editorial Pearson Prentice Hall. España 2004.

**Hernández, Paola Andrea Cardona.** *Aplicación de árboles de decisión en modelos de riesgo crediticio*. Revista colombiana de estadística Nº 2 Vol. 27. 2004. Disponible en: *http://www.emis.ams.org/journals/RCE/V27/V27_2_139Cardona.pdf*

**Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones (ICEA)**. *Caída en el Ramo de Vida. Octubre* 2013. Estadística Año 2012. Informe n° 1309. Disponible en: *www.icea.es*

**Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones (ICEA)**. *Caída en el Ramo de Vida. Octubre* 2014. Estadística Año 2013. Informe n° 1352. Disponible en: *www.icea.es***Jurado Gil, José.** *El Seguro de Vida en España: Factores que influyen en su progreso*. 2009. Fundación Mapfre. Disponible en: *www.fundacionmapfre.com/cienciasdelseguro/*

**Jackson, Donald.** *Determining a customer's lifetime value.* Direct Marketing Nº 51.11. Año 1989.

**Kaas R., Goovaerts M., Dhaene J. y Denuit M.** *Modern actuarial risk theory,* Kluwer Academic Publishers. Boston 2001.

**Kecman, Vojislav.** *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT press. Año 2001.

**Larose Daniel T.** *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Inc. USA 2005.

**Lindsey, James K.** *Applying generalized linear models.* Springer Science & Business Media. Año 1997.

**Martínez Campos, Francisco.** *Análisis de los Patrones de Conducta en la Fuga de Clientes mediante Técnicas de Inteligencia Artificial. Aplicación Práctica al Ramo de Decesos*. Universidad Complutense de Madrid. Septiembre 2014.

**McCullagh, Peter y John A. Nelder.** *Generalized linear models.* Vol. 37. Chapman & Hall Press. 2nd Edition. 1989.

**Mena, Jesús.** *Data mining your website*. Digital Press, Año 1999.

**Mena, Jesús**. *Machine-learning the business: Using data mining for competitive intelligence*. Competitive Intelligence Review Nº 4, Vol 7. Año 1996.

**Michalski, Ryszard.** *S. A theory and methodology of inductive learning*. Springer Berlín Heidelberg. Año 1983.

**Millán Aguilar, Adolfo y Muñoz Colomina C.I.** *Indicadores de calidad en el sector asegurador.* Cruzando fronteras: tendencias de contabilidad directiva para el siglo XXI: actas VII Congreso Internacional de Costos y II Congreso de la Asociación Española de Contabilidad Directiva. Servicio de Publicaciones, 2001. Disponible en: http://www.intercostos.org/documentos/Trabajo156.pdf

**Millán Aguilar, Adolfo y Muñoz Colomina C.I.** *Indicadores de calidad en el sector asegurador.* Cuadernos de Estudios Empresariales N° 10. Madrid 2010.

**Minsky, Marvin L.** *Computation: finite and infinite machines.* Prentice-Hall, Inc.. Año 1967.

**Miranda, M., Segovia, M., Gómez, P. y Blanco, S.** *Capítulo 7 La influencia del capital humano de las empresas industriales españolas en su intensidad exportadora: Análisis mediante la técnica PART de Inteligencia Artificial.* Estudios en Finanzas y Contabilidad: España y América Latina. Estado del arte y las nuevas metodologías aplicadas. Universidad Complutense de Madrid. Tópicos Selectos de Finanzas Vol° 1. Año 2013.

**Moscarola, Jean.** *Multicriteria Decision Aid Two Applications in Education Management*. Springer Berlin Heidelberg. Año 1978.

**Nelder, John A. y Baker, R. J.** *Generalized linear models.* Encyclopedia of Statistical Sciences. Año 1972.

**Nelder, J.A. y Wedderburn, R.W.M.** *Generalized Linear Models*. Journal of the Royal Statistical Society Vol° 135.3. Año 1972.

**Nelder, J.A. y McCullagh, P.** *Generalized Linear Models*. Segunda Edición. Chapman&Hall. Año 1989.

**Nurmi, Hannu, Janusz Kacprzyk, and Mario Fedrizzi.** *Probabilistic, fuzzy and rough concepts in social choice.* European Journal of Operational Research Volº 95.2. Año 1996.

**O'Leary, Daniel E.** *Using neural networks to predict corporate failure*. International Journal of Intelligent Systems in Accounting, Finance and Management. Disponible en: *https://msbfile03.usc.edu/digitalmeasures/doleary/intellcont/nncorporate%20failure-1.pdf*

**Ohlsson, E. y Johansson, B.** *Non-Life Insurance Pricing with Generalized Linear Models.* Heidelberg: SPRINGER. Año 2010.

**Pawlak, Zdzislaw.** *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers. Dordrecht, Boston, London 1991.

**Pawlak, Zdzislaw**. *Rough sets and intelligent data analysis*. Information Sciences. Año 2002.

Pawlak, z. y Skowron, A. *Rudiments of rough sets*. Decision Sciences N° 117. Año 2007.

**Pieschacón JR., C.A.** *La caída de cartera en vida y sus consecuencias*. Actualidad Aseguradora Nº 10. Año 2010.

**Predki, B., Slowinski, R., Stefanowski, J., Susmaga, R. y Wilk, S.** *ROSE – Software Implementation of the Rough Set Theory*. En POLKOWSKI, L. y SKOWRON, A. (Eds.): Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence Volº 1424. Springer-Verlag. Berlin 1998.

**Predki, B. y Wilk, S.** *Rough Set Based Data Exploration Using ROSE System.* En RAS, Z.W. y SKOWRON, A. (Eds.): Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence Volº 1609. Springer-Verlag. Berlin 1999.

**Puche, José Gabriel.** *Solvencia II. El riesgo de falta de armonización entre países*. Primavera 2013. Revista de Actuarios n° 32 del Instituto de Actuarios Españoles. Disponible en: *www.actuarios.org*

**Quinlan, J.R.** *Induction of Decision Trees*. Agosto 1985. Revista Machine Learning Volº 1-1. Disponible en: *http://link.springer.com/article/10.1023/A:1022643204877*

**Quinlan, J.R.** *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, Inc. California 1993.

**Reza, Fazlollah M**. *An introduction to information theory*. Courier Corporation. Año 1961.

**Rocco Roberto Cerchiara, Edwards Matthew y Gambini Alessandra**. *Generalized Linear Models in Life Insurance: Decrements and Risk Factor Analysis under Solvency II*. Universidad de Calabria, Italia.

**Rodríguez-Pardo del Castillo, José Miguel.** *Modelos predictivos aplicados al seguro de Vida*. Septiembre-Diciembre 2012. Revista de Gerencia de Riesgos y Seguros n° 114 de la Fundación Mapfre. Disponible en: *www.mapfre.com/fundacion/html/revistas/*

**Roubens, Marc y Vincke, Philippe**. *Preference modelling*. Lectures Notes in Economics and Mathematical Systems Volº 250. Año 1985.

**Roy, Richard y Kailath, Thomas**. *ESPRIT-estimation of signal parameters via rotational invariance techniques.* Acoustics, Speech and Signal Processing, IEEE Transactions Volº 37.7. Año 1989.

**Sáez de Jáuregui, Luis María.** *Solvencia II: Una realidad que obligará a hacer fácil lo difícil y con talento*. Primavera 2013. Revista de Actuarios n° 32 del Instituto de Actuarios Españoles. Disponible en: *www.actuarios.org*

**Sanchis, Alicia**. *Una aplicación del Análisis Discriminante a la previsión de la Insolvencia en las empresas españolas de seguros no-vida.* Tesis Doctoral, Universidad Complutense de Madrid. Año 2000

**Sanchis Arellano, A.; Gil Fana, J.A. y Heras Martínez, A.** *El análisis discriminante en la previsión de la insolvencia en las empresas de seguros de no vida*. Revista Española de Financiación y Contabilidad N° 32.116. Enero - Marzo 2003.

**Sanchis, A., Segovia, M. J., Gil, J. A., Heras, A., y Vilar, J. L.** *Rough sets and the role of the monetary policy in financial stability (macroeconomic problem) and the prediction of insolvency in insurance sector (microeconomic problem).* European Journal of Operational Research Nº 181(3). Año 2007.

**Schlesinger, Harris y von der Schulenburg, J-Matthias Graf**. *Consumer information and decisions to switch insurers.* Journal of Risk and Insurance. Año1993.

**Segovia Vargas, María Jesús.** *Predicción de Crisis Empresarial en Seguros No Vida mediante la Metodología Rough Set* (Tesis Doctoral). Universidad Complutense de Madrid. Madrid, 2003. Disponible en: *http://eprints.ucm.es/tesis/cee/ucm-t26780.pdf*

**Segovia Vargas, M.J., Fana, J. G., Martínez, A. H., Zanón, J. V., & Contabilidad, I.** *La metodología Rough Set frente al Análisis Discriminante en los problemas de clasificación multiatributo*. Universidad Complutense de Madrid. Oviedo 2003. Disponible en: *http://www.uv.es/sala/malaga/XI/19.pdf*

**Segovia Vargas, María Jesús, Miranda García, Marta y Escamilla Ramos, María.** *Técnicas de inteligencia artificial aplicadas a la resolución de problemas económico-financieros: análisis de los factores determinantes del éxito exportador.* Gestión Informática Empresarial. CES Felipe II, Universidad Complutense de Madrid y Universidad Tecnológica de México

**Shyng, Jhieh-Yu, Wang, F. K., Tzeng, G. H., y Wu, K. S.** *Rough set theory in analyzing the attributes of combination values for the insurance market*. Expert Systems with Applications Nº 32.1. Año 2007.

**Skowron, A. y Grzyma la-Busse.** *From the Rough Set Theory to the Evidence Theory*. Institute of Computer Science Reports 8. Año 1991

**Slowiński, Roman.** *Rough set learning of preferential attitude in multi-criteria decision making*. Methodologies for Intelligent Systems. Springer Berlin Heidelberg. Año 1993.

**Słowiński, Roman y Jerzy Stefanowski.** *Rough classification with valued closeness relation*. New approaches in classification and data analysis. Springer Berlin Heidelberg, Año 1994.

**Soley, Jorge.** *Solvencia II, Nota Técnica*. Dirección del Área de Solvencia II de Vida Caixa Grupo. Disponible en: *http://www.iese.edu/Aplicaciones/upload/SolvenciaIINotaTcnica2.pdf*

**Swiss Reinsurance Company Ltd. (Economic Research & Consulting).** *Informe SIGMA: El Seguro Mundial en 2014: Vuelta a la Vida.* Informe N° 4/2015. Disponible en: *http://www.swissre.com/library/#inline*

**Tolmos Rodríguez-Piñero, Piedad.** *SVM para la clasificación de asegurados en el seguro del automóvil.* Empresa global y mercados locales: XXI Congreso Anual AEDEM, Universidad Rey Juan Carlos. Escuela Superior de Gestión Comercial y Marketing, ESIC, Madrid 2007.

**Tolmos Rodrıguez-Piñero, Piedad y Mozos, R. S.** *Prediction of claims and risk factor selection in automobile insurance using Support Vector Machines and Genetic Algorithms*. New Trends and Tools in Complex Networks Nº 115. Madrid 2007.

**Unión Española de Entidades Aseguradoras y Reaseguradoras (UNESPA)**. *Memoria Social del Seguro Español.* Año 2012. Disponible en: *www.unespa.es*

**Unión Española de Entidades Aseguradoras y Reaseguradoras (UNESPA)**. *Memoria Social del Seguro Español.* Año 2013. Disponible en: *www.unespa.es*

**Unión Española de Entidades Aseguradoras y Reaseguradoras (UNESPA)**. *Memoria Social del Seguro Español.* Año 2013. Disponible en: *www.unespa.es*

**Wedderburn, Robert WM.** *Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method.* Biometrika 61.3. Año 1974.

**Wilson, S. & Press, S. J.** *Choosing between logistic regression and discriminant analysis.* American Statistical Association 73.

**Witten, Ian H. y Eibe Frank.** *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. 2nd Edition. San Francisco 2005.

**Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., y Cunningham, S. J.** Weka: *Practical machine learning tools and techniques with Java implementations*. Año 1999.

**INDICE DE FIGURAS**

Figura 1. Esquema Conceptual del Proyecto Solvencia II ..... 29

Figura 2. Cálculo del SCR (Requerimiento de Capital de Solvencia) bajo la Fórmula Estándar..... 33

Figura 3. Gráfico de la Evolución de la Retención de Cartera ..... 45

Figura 4. Distribución de Caída de Cartera por Causas..... 48

Figura 5. Crecimiento de Primas del Ramo de Vida antes y después de la crisis económica del 2008 ..... 51

Figura 6. Densidad y penetración del seguro en los mercados avanzados en el 2014 .. 52

Figura 7. Histograma - EDAD ..... 53

Figura 8. Gráfico de la Distribución por SEXO ..... 58

Figura 9. Perfil de Fallecimientos por Sexo ..... 59

Figura 10. Esfuerzo de los hogares por adquirir seguros, según la edad de su sustentador principal ..... 60

Figura 11. Gráfico de la Distribución por EDAD..... 60

Figura 12. Gráfico de la Distribución por ANTIGÜEDAD ..... 62

Figura 13. Gráfico de la Distribución por TIPO DE PRODUCTO ..... 63

Figura 14. Gráfico de la Distribución por TIPO DE PRIMA..... 65

Figura 15. Gráfico de la Distribución por RED ..... 67

Figura 16. Gráfico de la Distribución por FORMA DE PAGO ..... 69

Figura 17. Gráfico de la Distribución por AÑO EFECTO ..... 70

Figura 18. Gráfico de la Distribución por ESTADO CIVIL ..... 71

Figura 19. Gráfico de la Distribución por HIJOS..... 73

Figura 20. Gráfico de la Distribución por VALOR DEL CLIENTE ..... 74

Figura 21. Gráfico de la Distribución por ICE..... 75

Figura 22. Gráfico de la Distribución por NIVEL DE INGRESOS ..... 76

Figura 23. Gráfico de la Distribución por NIVEL DE ESTUDIOS..... 78

Figura 24. Ejemplo de Árbol. De Decisión ..... 91

Figura 25. Salida de Resultados – WEKA (Algoritmo C4.5) ..... 106

Figura 26. Salida de Resultados del Árbol de Decisión – WEKA (Algoritmo C4.5) ..... 108

Figura 27. Composición de la cartera para la subdivisión de los análisis de acuerdo a la variable ANTIGÜEDAD..... 110

Figura 28. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 1) _ Regla 1 de la CLASE 1 ... 112

Figura 29. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 1) _ Regla 2 de la CLASE 1 ... 113

Figura 30. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 2) _ Regla 3 de la CLASE 1 ... 114

Figura 31. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 1) _ Regla 1 de la CLASE 0 ... 117

Figura 32. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 2) _ Regla 2 de la CLASE 0 ... 118

Figura 33. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 3) _ Regla 3 de la CLASE 0 ... 119

Figura 34. Árbol. De Decisión (Ramo de ANTIGÜEDAD = 3) _ Regla 4 de la CLASE 0 ... 120

Figura 35. Salida de Validación Cruzada (Rough Set) ..... 130

Figura 36. Reglas con Mayor Fuerza – Rough Set (CATEGORIA 1=Cancelación)..... 132

Figura 37. Reglas con Mayor Fuerza – Rough Set (CATEGORIA 0=Retención)..... 133

Figura 38. Traducción de Modelo Predictivo en Seguros de Vida..... 142

Figura 39. Ejemplo de “Estimación de un Factor” junto con sus “Errores Estándar” .. 167

Figura 40. Ejemplo de “Estimación de un Factor” junto con sus “Errores Estándar” con poca significancia ..... 168

Figura 41. Gráfico de Residuos con un comportamiento constante ..... 171

Figura 42. Gráfico de Residuos con un comportamiento irregular ..... 171

Figura 43. Gráfico de Leverage identificando valores atípicos ..... 172

Figura 44. Gráfico de Transformación de Box-Cox para los resultados de un Modelo de frecuencias ..... 173

Figura 45. Gráfico de Transformación de Box-Cox para los resultados de un Modelo de Severidad ..... 174

Figura 46. Gráfico de Impacto de la variable Suma Asegurada ..... 176

Figura 47. Gráfico de Impacto de la variable Año Calendario ..... 177

Figura 48. Salida de Resultados Modelo I – R (GLM, Binomial, Logit) ..... 192

Figura 49. Salida de Resultados Modelo II – R (GLM, Binomial, Logit) ..... 193

Figura 50. Salida de Resultados Modelo III – R (GLM, Binomial, Logit) ..... 194

Figura 51. Salida de Resultados Modelo IV – R (GLM, Binomial, Logit) ..... 195

Figura 52. Salida de Resultados Modelo V – R (GLM, Binomial, Logit) ..... 196

Figura 53. Salida de Resultados Modelo VI – R (GLM, Binomial, Logit) ..... 197

Figura 54. Salida de Resultados ANOVA Modelo V ..... 198

Figura 55. Salida de Resultados ANOVA Modelo VI ..... 198

Figura 56. Diferentes tipos de Curvas ROC ..... 200

Figura 57. Curva de ROC del Modelo VI – R (GLM, Binomial, Logit) ..... 200

Figura 58. Efecto de la variable FORMA PAGO con respecto a su Nivel Base ..... 201

Figura 59. Probabilidad de Caída o Abandono vs Antigüedad de la Póliza ..... 202

Figura 60. Efecto de la variable EDAD dentro de un GLM para el Ramo de No Vida ... 203

Figura 61. Efecto de la variable DURACION dentro de un GLM para el Ramo de No Vida ..... 204

Figura 62. Efecto de la variable PRODUCTO dentro de un GLM para el Ramo de No Vida ..... 204

**INDICE DE TABLAS**

Tabla 1. Histórico de Vida Media de la Cartera de Seguros de Vida Individual ..... 45

Tabla 2. Tasas de Caída de Cartera por Tipo de Producto al cierre del 2012 ..... 46

Tabla 3. Variación de Primas en el año 2014 ..... 51

Tabla 4. Estadística general de la cartera muestra - Edad ..... 53

Tabla 5. Variables seleccionadas para la aplicación empírica ..... 57

Tabla 6. Distribución de la muestra por la variable SEXO ..... 58

Tabla 7. Rangos de Edad de agrupación de la muestra ..... 60

Tabla 8. Distribución de la muestra por la variable ANTIGÜEDAD ..... 62

Tabla 9. Distribución de la muestra por la variable TIPO DE PRODUCTO ..... 63

Tabla 10. Modalidades del Seguro de Vida en España en el 2007 ..... 64

Tabla 11. Distribución de la muestra por la variable TIPO DE PRIMA ..... 65

Tabla 12. Distribución de la muestra por la variable RED ..... 67

Tabla 13. Distribución del Seguro de Vida por Canales ..... 67

Tabla 14. Distribución de la muestra por la variable FORMA DE PAGO ..... 68

Tabla 15. Distribución de la muestra por la variable AÑO EFECTO ..... 70

Tabla 16. Distribución de la muestra por la variable ESTADO CIVIL ..... 71

Tabla 17. Tasas de Penetración según el estado civil ..... 72

Tabla 18. Distribución de la muestra por la variable HIJOS ..... 73

Tabla 19. Distribución de la muestra por la variable VALOR DEL CLIENTE ..... 74

Tabla 20. Distribución de la muestra por la variable ICE ..... 75

Tabla 21. Distribución de la muestra por la variable NIVEL DE INGRESOS ..... 76

Tabla 22. Frecuencias de hogares que gastan en seguros según nivel de ingresos..... 77

Tabla 23. Distribución de la muestra por la variable NIVEL DE ESTUDIOS ..... 78

Tabla 24. Tasas de Penetración según el nivel de estudios ..... 79

Tabla 25. Tabla de Decisión - Ejemplo ..... 100

Tabla 26. Matriz de Diferenciación - Ejemplo ..... 101

Tabla 27. Distribución de acuerdo a la variable TIPO DE PRESTACION ..... 105

Tabla 28. Resumen de Resultados Arboles de Decisión – CLASE 1: CAIDA..... 124

Tabla 29. Resumen de Resultados Arboles de Decisión – CLASE 0: RETENCION ..... 127

Tabla 30. Distribución de acuerdo a la variable TIPO DE PRESTACION ..... 129

Tabla 31. Resumen de Resultados Rough Set – AMBAS CATEGORIAS: 1-CAIDA y 0-RETENCION ..... 135

Tabla 32. Distribuciones de la Familia Exponencial (parámetros y función de varianza).. 149

Tabla 33. Funciones Vínculo ..... 150

Tabla 34. Estructuras de Modelos más comunes ..... 153

Tabla 35. Funciones de Devianza..... 155

Tabla 36. Autovalores (Análisis de Componentes Principales) ..... 189

Tabla 37. Autovectores de los Componentes Principales ..... 190

Tabla 38. Comparativo de Resultados de Metodologías – CLASE 1: CAIDA ..... 205

Tabla 39. Comparativo de Resultados de Metodologías – CLASE 0: RETENCION ..... 206